THE UNIVERSITY *of* EDINBURGH
informatics

Applied Machine Learning (AML)

Evaluation

Oisin Mac Aodha • Siddharth N.

Evaluation

Outline

Evaluation Measures

- How (in)accurate is a model?
- Supervised Learning
  - Classification
  - Regression
- Unsupervised Learning

Outline

Evaluation Measures

- How (in)accurate is a model?
- Supervised Learning
  - Classification
  - Regression
- Unsupervised Learning

Evaluation

Classification

Classification

Naive Bayes: Spam

email	true
“send us your password”	+
“send us review”	−
“review your account”	−
“review us”	+
“send your password”	+
“send us your account”	+
⋮	

Classification

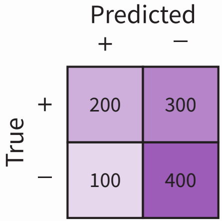
Naive Bayes: Spam

email	true	pred
“send us your password”	+	+
“send us review”	−	+
“review your account”	−	−
“review us”	+	−
“send your password”	+	+
“send us your account”	+	+
⋮		

Classification

Naive Bayes: Spam

email	true	pred
“send us your password”	+	+
“send us review”	−	+
“review your account”	−	−
“review us”	+	−
“send your password”	+	+
“send us your account”	+	+
⋮		



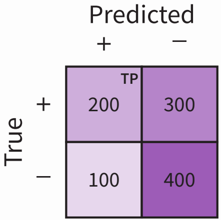
		Predicted	
		+	−
True	+	200	300
	−	100	400

Classification

Naive Bayes: Spam

email	true	pred
“send us your password”	+	+
“send us review”	−	+
“review your account”	−	−
“review us”	+	−
“send your password”	+	+
“send us your account”	+	+

⋮

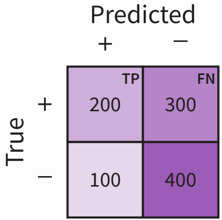


		Predicted	
		+	−
True	+	TP 200	300
	−	100	400

Classification

Naive Bayes: Spam

email	true	pred
“send us your password”	+	+
“send us review”	−	+
“review your account”	−	−
“review us”	+	−
“send your password”	+	+
“send us your account”	+	+
⋮		

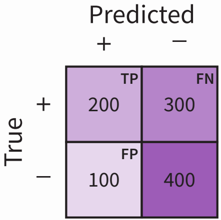


		Predicted	
		+	−
True	+	TP 200	FN 300
	−	100	400

Classification

Naive Bayes: Spam

email	true	pred
“send us your password”	+	+
“send us review”	−	+
“review your account”	−	−
“review us”	+	−
“send your password”	+	+
“send us your account”	+	+
⋮		

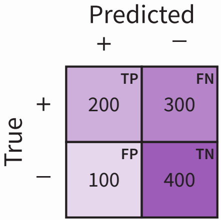


		Predicted	
		+	−
True	+	TP 200	FN 300
	−	FP 100	400

Classification

Naive Bayes: Spam

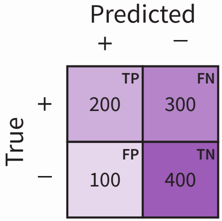
email	true	pred
“send us your password”	+	+
“send us review”	−	+
“review your account”	−	−
“review us”	+	−
“send your password”	+	+
“send us your account”	+	+
⋮		



		Predicted	
		+	−
True	+	TP 200	FN 300
	−	FP 100	TN 400

Classification Error

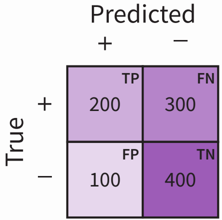
- $\text{Error} = \dfrac{\text{incorrect}}{\text{total}} = \dfrac{\mathrm{FP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$



		Predicted	
		+	−
True	+	TP 200	FN 300
	−	FP 100	TN 400

Classification Error

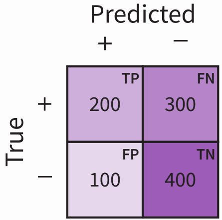
- $\text{Error} = \dfrac{\text{incorrect}}{\text{total}} = \dfrac{\mathrm{FP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$
- $\text{Accuracy} = \dfrac{\text{correct}}{\text{total}} = \dfrac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$



		Predicted	
		+	−
True	+	**TP** 200	**FN** 300
	−	**FP** 100	**TN** 400

Classification Error

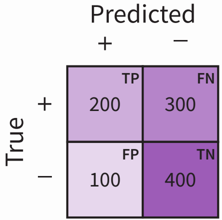
- $\text{Error} = \dfrac{\text{incorrect}}{\text{total}} = \dfrac{\mathrm{FP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$
- $\text{Accuracy} = \dfrac{\text{correct}}{\text{total}} = \dfrac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$



		Predicted	
		+	−
True	+	**TP** 200	**FN** 300
	−	**FP** 100	**TN** 400

Classification Error

- $\text{Error} = \dfrac{\text{incorrect}}{\text{total}} = \dfrac{\mathrm{FP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$
- $\text{Accuracy} = \dfrac{\text{correct}}{\text{total}} = \dfrac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$

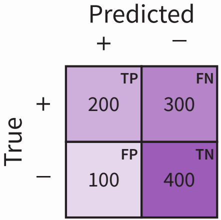


		Predicted	
		+	−
True	+	**TP** 200	**FN** 300
	−	**FP** 100	**TN** 400

Measure of how “good” a classifier is

Classification Error

- $\text{Error} = \frac{\text{incorrect}}{\text{total}} = \frac{\mathrm{FP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$
- $\text{Accuracy} = \frac{\text{correct}}{\text{total}} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$

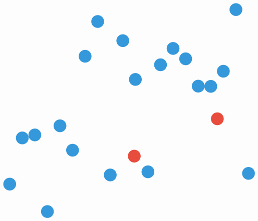


		Predicted	
		+	−
True	+	TP 200	FN 300
	−	FP 100	TN 400

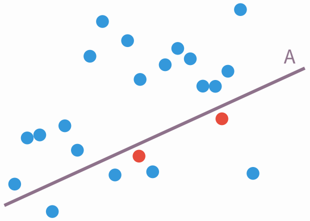
Measure of how “good” a classifier is

Issue: Class imbalances

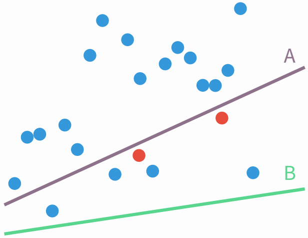
Classification Accuracy: Class Imbalance



Classification Accuracy: Class Imbalance

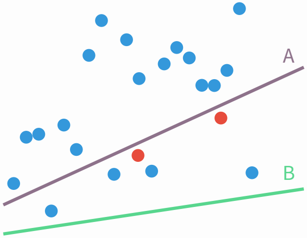


Classification Accuracy: Class Imbalance



Acc(B) > Acc(A)!

Classification Accuracy: Class Imbalance

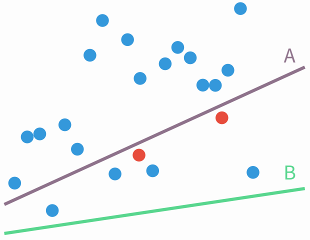


Acc(B) > Acc(A)!

Examples

- Earthquakes: rare event
  → very good accuracy if always '–' !

Classification Accuracy: Class Imbalance

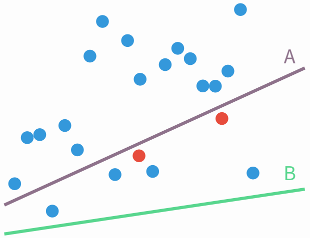


Acc(B) > Acc(A)!

Examples

- Earthquakes: rare event
  → very good accuracy if always '–' !
- Web search: mostly irrelevant
  → very good accuracy if always "irrelevant"!

Classification Accuracy: Class Imbalance

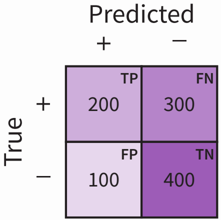


Acc(B) > Acc(A)!

Examples

- Earthquakes: rare event
  → very good accuracy if always '–' !
- Web search: mostly irrelevant
  → very good accuracy if always "irrelevant"!
- Cost of errors (FN, FP) are not the same

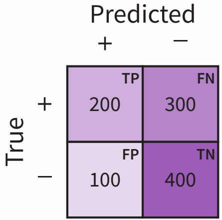
Error Measures



		Predicted	
		+	−
True	+	**TP** 200	**FN** 300
	−	**FP** 100	**TN** 400

Error Measures

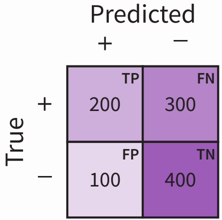
- False Positive Rate (FPR) = $\dfrac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$
    - (False Alarm) % of ‘–’ misclassified as ‘+’



		Predicted	
		+	–
True	+	TP 200	FN 300
	–	FP 100	TN 400

Error Measures

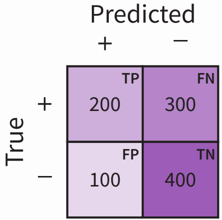
- False Positive Rate (FPR) = $\dfrac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}$
    - (False Alarm) % of ‘–’ misclassified as ‘+’
- False Negative Rate (FNR) = $\dfrac{\mathrm{FN}}{\mathrm{TP}+\mathrm{FN}}$
    - (Miss) % of ‘+’ misclassified as ‘–’



		Predicted	
		+	–
True	+	TP 200	FN 300
	–	FP 100	TN 400

Error Measures

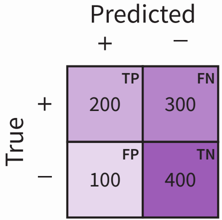
- False Positive Rate (FPR) = $\dfrac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$
    - (False Alarm) % of ‘–’ misclassified as ‘+’
- False Negative Rate (FNR) = $\dfrac{\mathrm{FN}}{\mathrm{TP} + \mathrm{FN}}$
    - (Miss) % of ‘+’ misclassified as ‘–’
- Recall / True Positive Rate (TPR) = $\dfrac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$
    - (1 - Miss) % of ‘+’ correctly predicted



		Predicted	
		+	–
True	+	TP 200	FN 300
	–	FP 100	TN 400

Error Measures

- False Positive Rate (FPR) = $\dfrac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$
  - (False Alarm) % of ‘–’ misclassified as ‘+’
- False Negative Rate (FNR) = $\dfrac{\mathrm{FN}}{\mathrm{TP} + \mathrm{FN}}$
  - (Miss) % of ‘+’ misclassified as ‘–’
- Recall / True Positive Rate (TPR) = $\dfrac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$
  - (1 - Miss) % of ‘+’ correctly predicted
- Precision / Positive Predictive Rate (PPR) = $\dfrac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$
  - % of ‘+’ out of all positive predictions

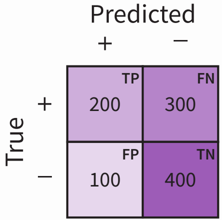


		Predicted	
		+	–
True	+	TP 200	FN 300
	–	FP 100	TN 400

Error Measures

- False Positive Rate (FPR) = $\dfrac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$
  - (False Alarm) % of ‘–’ misclassified as ‘+’
- False Negative Rate (FNR) = $\dfrac{\mathrm{FN}}{\mathrm{TP} + \mathrm{FN}}$
  - (Miss) % of ‘+’ misclassified as ‘–’
- Recall / True Positive Rate (TPR) = $\dfrac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$
  - (1 - Miss) % of ‘+’ correctly predicted
- Precision / Positive Predictive Rate (PPR) = $\dfrac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$
  - % of ‘+’ out of all positive predictions

Report pairs: Precision—Recall; TPR—FPR (ROC)

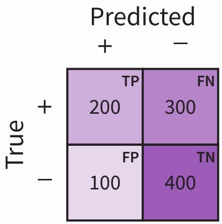


	Predicted +	Predicted −
True +	TP 200	FN 300
True −	FP 100	TN 400

Error Measures

Unified Measures:

optimisation objective, comparative evaluation



		Predicted	
		+	−
True	+	**TP** 200	**FN** 300
	−	**FP** 100	**TN** 400

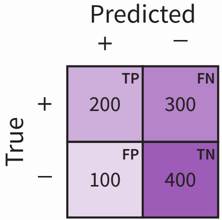
Error Measures

Unified Measures:

optimisation objective, comparative evaluation

- Detection cost

  cost = $C_{\mathrm{FP}} \cdot FP + C_{\mathrm{FN}} \cdot FN$



		Predicted	
		+	−
True	+	**TP** 200	**FN** 300
	−	**FP** 100	**TN** 400

Error Measures

Unified Measures:

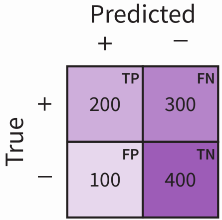
optimisation objective, comparative evaluation

- Detection cost

  cost = $C_{\mathrm{FP}} \cdot FP + C_{\mathrm{FN}} \cdot FN$

- F-measure

  harmonic mean of precision (Pr) & recall (Re): $F1 = \dfrac{2 \cdot \mathrm{Pr} \cdot \mathrm{Re}}{\mathrm{Pr} + \mathrm{Re}}$

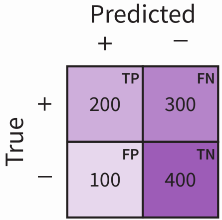


		Predicted	
		+	−
True	+	TP 200	FN 300
	−	FP 100	TN 400

Error Measures

Unified Measures:

optimisation objective, comparative evaluation



		Predicted	
		+	−
True	+	TP 200	FN 300
	−	FP 100	TN 400

- Detection cost

  cost = $C_{\mathrm{FP}} \cdot FP + C_{\mathrm{FN}} \cdot FN$

- F-measure

  harmonic mean of precision (Pr) & recall (Re): $F1 = \dfrac{2 \cdot \mathrm{Pr} \cdot \mathrm{Re}}{\mathrm{Pr} + \mathrm{Re}}$

- Cohen's Kappa $\kappa = \frac{p_o - p_e}{1 - p_e}$

Error Measures

Unified Measures:

optimisation objective, comparative evaluation

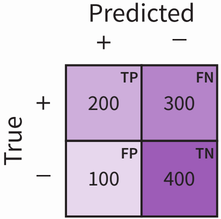
- Detection cost

  cost = $C_{\mathrm{FP}} \cdot FP + C_{\mathrm{FN}} \cdot FN$

- F-measure

  harmonic mean of precision (Pr) & recall (Re): $F1 = \dfrac{2 \cdot \mathrm{Pr} \cdot \mathrm{Re}}{\mathrm{Pr} + \mathrm{Re}}$

- Cohen's Kappa $\kappa = \frac{p_o - p_e}{1 - p_e}$
  - p_o = label agreement b/w model predictions and targets (accuracy)

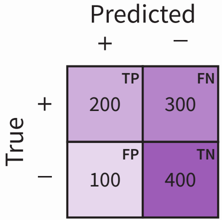


		Predicted	
		+	−
True	+	TP 200	FN 300
	−	FP 100	TN 400

Error Measures

Unified Measures:

optimisation objective, comparative evaluation



		Predicted	
		+	−
True	+	TP 200	FN 300
	−	FP 100	TN 400

- Detection cost

  cost = $C_{\mathrm{FP}} \cdot FP + C_{\mathrm{FN}} \cdot FN$

- F-measure

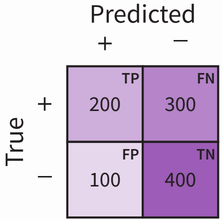
  harmonic mean of precision (Pr) & recall (Re): $F1 = \dfrac{2 \cdot \mathrm{Pr} \cdot \mathrm{Re}}{\mathrm{Pr} + \mathrm{Re}}$

- Cohen's Kappa $\kappa = \frac{p_o - p_e}{1 - p_e}$
  - p_o = label agreement b/w model predictions and targets (accuracy)
  - p_e = chance agreement b/w model predictions and targets

Error Measures

Unified Measures:

optimisation objective, comparative evaluation



		Predicted	
		+	−
True	+	TP 200	FN 300
	−	FP 100	TN 400

- Detection cost

  cost = $C_{\mathrm{FP}} \cdot FP + C_{\mathrm{FN}} \cdot FN$

- F-measure

  harmonic mean of precision (Pr) & recall (Re): $F1 = \dfrac{2 \cdot \mathrm{Pr} \cdot \mathrm{Re}}{\mathrm{Pr} + \mathrm{Re}}$

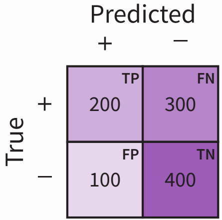
- Cohen's Kappa $\kappa = \frac{p_o - p_e}{1 - p_e}$
  - p_o = label agreement b/w model predictions and targets (accuracy)
  - p_e = chance agreement b/w model predictions and targets

    denoting total $\mathrm{T} = \mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}$

Error Measures

Unified Measures:

optimisation objective, comparative evaluation



		Predicted	
		+	−
True	+	TP 200	FN 300
	−	FP 100	TN 400

- Detection cost

  cost = $C_{\mathrm{FP}} \cdot FP + C_{\mathrm{FN}} \cdot FN$

- F-measure

  harmonic mean of precision (Pr) & recall (Re): $F1 = \dfrac{2 \cdot \mathrm{Pr} \cdot \mathrm{Re}}{\mathrm{Pr} + \mathrm{Re}}$

- Cohen's Kappa $\kappa = \frac{p_o - p_e}{1 - p_e}$
  - p_o = label agreement b/w model predictions and targets (accuracy)
  - p_e = chance agreement b/w model predictions and targets

    denoting total $\mathrm{T} = \mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}$

    $= \frac{\mathrm{TP+FP}}{\mathrm{T}} \cdot \frac{\mathrm{TP+FN}}{\mathrm{T}} + \frac{\mathrm{TN+FN}}{\mathrm{T}} \cdot \frac{\mathrm{TN+FP}}{\mathrm{T}}$

Thresholds

- Models typically compute “confidence” as $p(y|\boldsymbol{x})$

Thresholds

- Models typically compute “confidence” as $p(y|\boldsymbol{x})$
- Decisions are made by thresholding this confidence

  $p(y|\boldsymbol{x}) > \tau \implies \texttt{spam}$

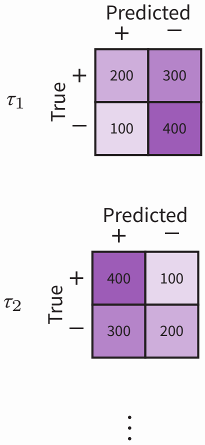
  - Logistic regression: $\sigma(\boldsymbol{w}^\top \boldsymbol{x}) \geq 0.5$
  - Naive Bayes: $p(y = \texttt{spam}|x) > 0.5$

Thresholds

- Models typically compute “confidence” as $p(y|\boldsymbol{x})$
- Decisions are made by thresholding this confidence

  $p(y|\boldsymbol{x}) > \tau \implies \texttt{spam}$
  - Logistic regression: $\sigma(\boldsymbol{w}^\top \boldsymbol{x}) \geq 0.5$
  - Naive Bayes: $p(y = \texttt{spam}|x) > 0.5$
- τ determines error rates and confusion matrix

  each τ provides a value for chosen measure(s)



τ_1

True \ Predicted	+	−
+	200	300
−	100	400

τ_2

True \ Predicted	+	−
+	400	100
−	300	200

⋮

Thresholds

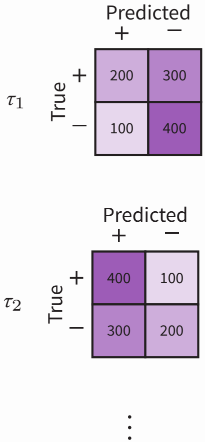
- Models typically compute “confidence” as $p(y|\boldsymbol{x})$
- Decisions are made by thresholding this confidence

  $p(y|\boldsymbol{x}) > \tau \implies \texttt{spam}$

  - Logistic regression: $\sigma(\boldsymbol{w}^\top \boldsymbol{x}) \geq 0.5$
  - Naive Bayes: $p(y = \texttt{spam}|x) > 0.5$
- τ determines error rates and confusion matrix

  each τ provides a value for chosen measure(s)
- **Complete picture:**

  plot measures as τ varies from $-\infty$ to ∞



τ_1

		Predicted	
		+	−
True	+	200	300
	−	100	400

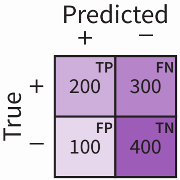
τ_2

		Predicted	
		+	−
True	+	400	100
	−	300	200

⋮

Precision-Recall Curve

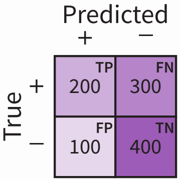
$$\text{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \qquad \text{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$



		Predicted	
		+	−
True	+	TP 200	FN 300
	−	FP 100	TN 400

Precision-Recall Curve

$$\text{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \qquad \text{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$



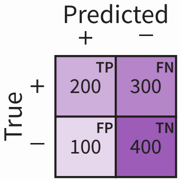
		Predicted	
		+	−
True	+	**TP** 200	**FN** 300
	−	**FP** 100	**TN** 400

email	label	$p(y\|\boldsymbol{x})$
“send us your password”	+	0.92
“send us review”	−	0.80
“review your account”	−	0.72
“review us”	+	0.65
“send your password”	+	0.61
“send us your account”	+	0.43

⋮

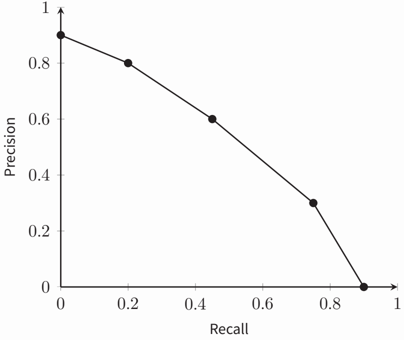
Precision-Recall Curve

$$\text{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \qquad \text{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$



		Predicted +	Predicted −
True	+	TP 200	FN 300
	−	FP 100	TN 400

email	label	$p(y \mid \boldsymbol{x})$
“send us your password”	+	0.92
“send us review”	−	0.80
“review your account”	−	0.72
“review us”	+	0.65
“send your password”	+	0.61
“send us your account”	+	0.43
⋮		

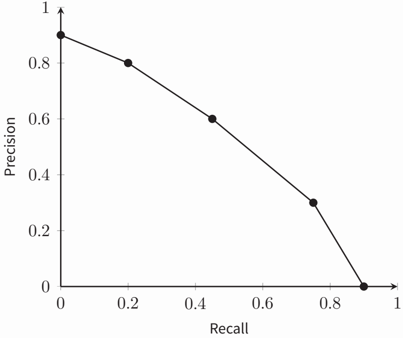


Precision-Recall Curve

$$\text{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \qquad \text{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

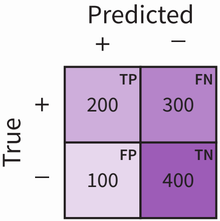
		Predicted	
		+	−
True	+	TP 200	FN 300
	−	FP 100	TN 400

email	label	$p(y\|\boldsymbol{x})$
“send us your password”	+	0.92
“send us review”	−	0.80
“review your account”	−	0.72
“review us”	+	0.65
“send your password”	+	0.61
“send us your account”	+	0.43
⋮		

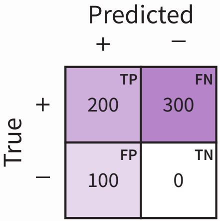


Area under curve (AUC) provides a more complete measure.

Precision-Recall: Issues



		Predicted	
		+	−
True	+	**TP** 200	**FN** 300
	−	**FP** 100	**TN** 400



		Predicted	
		+	−
True	+	**TP** 200	**FN** 300
	−	**FP** 100	**TN** 0

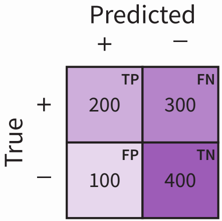
Precision-Recall: Issues

		Predicted	
		+	−
True	+	TP 200	FN 300
	−	FP 100	TN 400

		Predicted	
		+	−
True	+	TP 200	FN 300
	−	FP 100	TN 0

- Both classifiers get $\mathrm{Pr} = 66.7\%$ and $\mathrm{Re} = 40\%$

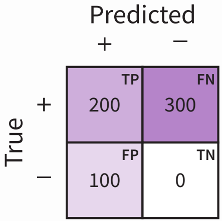
Precision-Recall: Issues



Predicted

True

	+	−
+	TP 200	FN 300
−	FP 100	TN 400



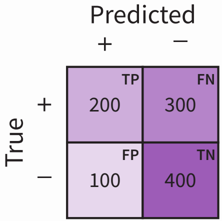
Predicted

True

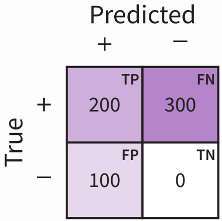
	+	−
+	TP 200	FN 300
−	FP 100	TN 0

- Both classifiers get $\mathrm{Pr} = 66.7\%$ and $\mathrm{Re} = 40\%$
  - Same positive recognition rate (66.7%)

Precision-Recall: Issues



		Predicted	
		+	−
True	+	TP 200	FN 300
	−	FP 100	TN 400



		Predicted	
		+	−
True	+	TP 200	FN 300
	−	FP 100	TN 0

- Both classifiers get $\mathrm{Pr} = 66.7\%$ and $\mathrm{Re} = 40\%$
  - Same positive recognition rate (66.7%)
  - *Very different* negative recognition rates: strong on left, nil on right

Receiver Operating Characteristic (ROC)

True Postive Rate (TPR) / Recall = $\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$ False Positive Rate (FPR) = $\frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$

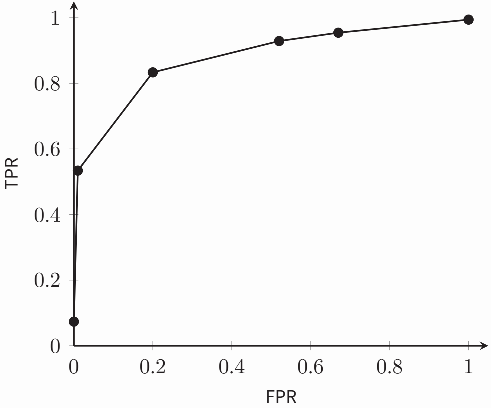
Receiver Operating Characteristic (ROC)

$$\text{True Postive Rate (TPR) / Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad \text{False Positive Rate (FPR)} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

AUC

- area under ROC curve
- larger area $\Longrightarrow$ better model

		Predicted	
		+	−
True	+	TP 200	FN 300
	−	FP 100	TN 400



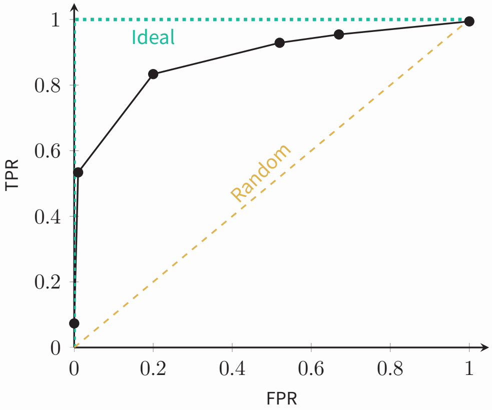
Receiver Operating Characteristic (ROC)

$$\text{True Postive Rate (TPR) / Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad \text{False Positive Rate (FPR)} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

AUC

- area under ROC curve
- larger area $\implies$ better model

		Predicted +	Predicted −
True	+	**TP** 200	**FN** 300
	−	**FP** 100	**TN** 400



Multi-Class Classification

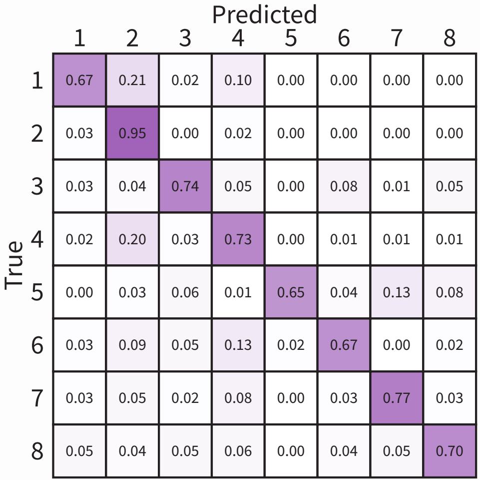
Measures

- Cohen’s Kappa $\kappa = \dfrac{p_o - p_e}{1 - p_e}$
- Matthews Correlation Coefficient (MCC)

$$\mathrm{MCC} = \frac{p_0 - p_e}{\sqrt{(1 - p_y)(1 - p_{\hat{y}})}}$$

$$p_e = \sum_{k=1}^{K} \left(\frac{1}{T} \sum_{j=1}^{K} C_{k,j} \right) \left(\frac{1}{T} \sum_{i=1}^{K} C_{i,k} \right) \qquad T = \sum_{i,j} C_{i,j}$$

$$p_y = \sum_{k=1}^{K} \left(\frac{1}{T} \sum_{j=1}^{K} C_{k,j} \right)^2 \qquad p_{\hat{y}} = \sum_{k=1}^{K} \left(\frac{1}{T} \sum_{i=1}^{K} C_{i,k} \right)^2$$



		Predicted							
		1	2	3	4	5	6	7	8
True	1	0.67	0.21	0.02	0.10	0.00	0.00	0.00	0.00
	2	0.03	0.95	0.00	0.02	0.00	0.00	0.00	0.00
	3	0.03	0.04	0.74	0.05	0.00	0.08	0.01	0.05
	4	0.02	0.20	0.03	0.73	0.00	0.01	0.01	0.01
	5	0.00	0.03	0.06	0.01	0.65	0.04	0.13	0.08
	6	0.03	0.09	0.05	0.13	0.02	0.67	0.00	0.02
	7	0.03	0.05	0.02	0.08	0.00	0.03	0.77	0.03
	8	0.05	0.04	0.05	0.06	0.00	0.04	0.05	0.70

Evaluation

Regression

Evaluating Regression

Classification

- *count* how often incorrect

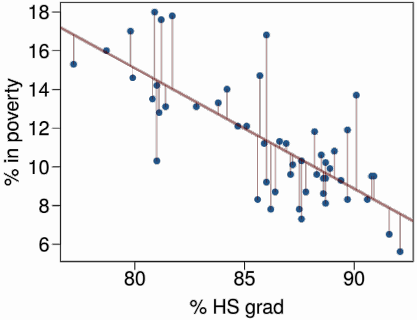
Evaluating Regression

Classification

- *count* how often incorrect

Regression

- *always* wrong (!) …but by how much?



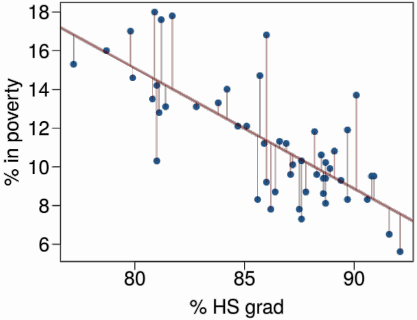
Evaluating Regression

Classification

- *count* how often incorrect

Regression

- *always* wrong (!) …but by how much?
- distance between predicted and actual values



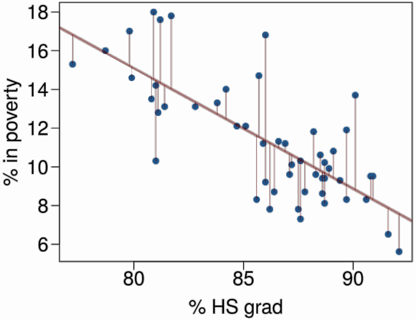
Evaluating Regression

Classification

- *count* how often incorrect

Regression

- *always* wrong (!) …but by how much?
- distance between predicted and actual values



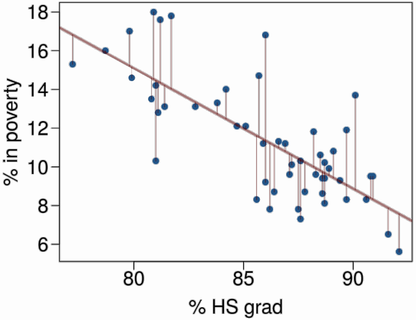
Evaluating Regression

Classification

- *count* how often incorrect

Regression

- *always* wrong (!) …but by how much?
- distance between predicted and actual values



$$\mathrm{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

Evaluating Regression

$$\mathrm{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

Evaluating Regression

$$\mathrm{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

Error Measures

- Mean Square Error (MSE)

$$\mathrm{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Evaluating Regression

$$\mathrm{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

Error Measures

- Mean Square Error (MSE)
- Mean Absolute Error (MAE)

$$\mathrm{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|$$

Evaluating Regression

$$\mathrm{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

Error Measures

- Mean Square Error (MSE)
- Mean Absolute Error (MAE)
- Correlation Coefficient (ρ)

$$\mathrm{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{\mathrm{Cov}(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\sqrt{\mathrm{Var}(\boldsymbol{y})}\sqrt{\mathrm{Var}(\hat{\boldsymbol{y}})}}$$

Evaluating Regression

$$\mathrm{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

Error Measures

- Mean Square Error (MSE)
- Mean Absolute Error (MAE)
- Correlation Coefficient (ρ)
- Coefficient of Determination (R^2)

$$\mathrm{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = 1 - \frac{\mathrm{MSE}(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\mathrm{Var}(\boldsymbol{y})}$$

Mean Squared Error

$$\mathrm{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Mean Squared Error

$$\mathrm{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Characteristics

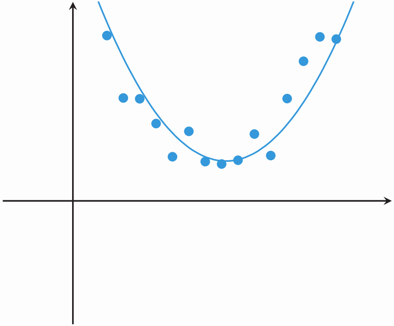
- Sensitivity to outliers
    - squaring → blow up error

Mean Squared Error

$$\text{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Characteristics

- Sensitivity to outliers
  - squaring → blow up error

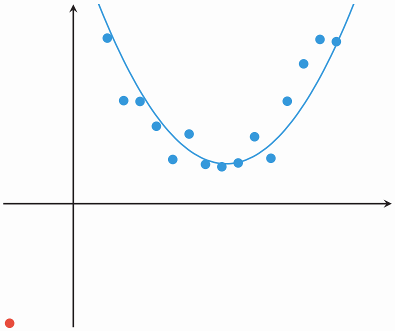


Mean Squared Error

$$\mathrm{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Characteristics

- Sensitivity to outliers
  - squaring → blow up error

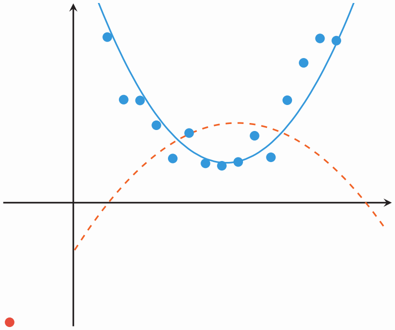


Mean Squared Error

$$\mathrm{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Characteristics

- Sensitivity to outliers
  - squaring → blow up error

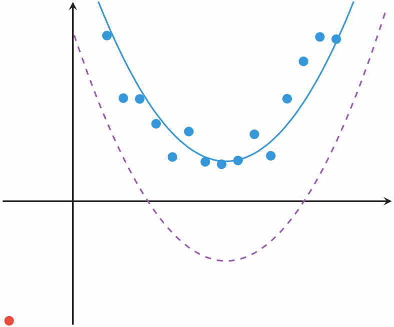


Mean Squared Error

$$\mathrm{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Characteristics

- Sensitivity to outliers
  - squaring → blow up error
- Sensitivity to scaling / translation

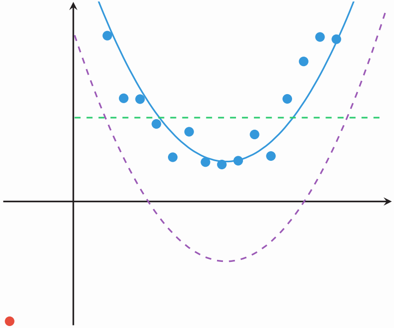


Mean Squared Error

$$\operatorname{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Characteristics

- Sensitivity to outliers
  - squaring → blow up error
- Sensitivity to scaling / translation
- **Baseline:** predict mean y



Mean Absolute Error

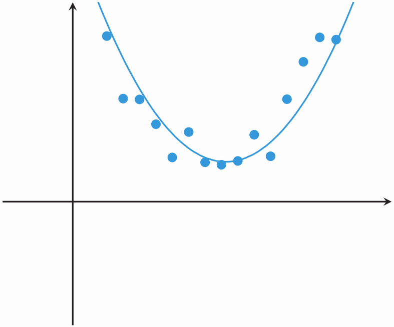
$$\mathrm{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|$$

Mean Absolute Error

$$\mathrm{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|$$

Characteristics

- Less sensitive to outliers
    - no squaring

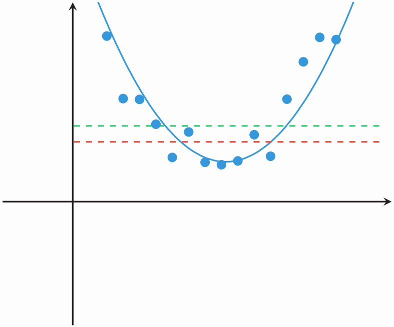


Mean Absolute Error

$$\text{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

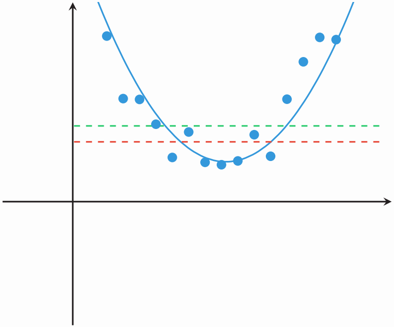
Characteristics

- Less sensitive to outliers
  - no squaring
- **Baseline:** median, not mean



Mean Absolute Error

$$\mathrm{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$



Characteristics

- Less sensitive to outliers
    - no squaring
- **Baseline:** median, not mean

Variants

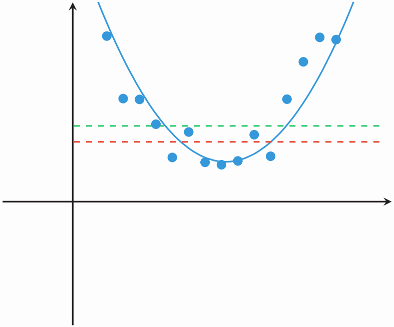
- Median Absolute Error: $= \mathrm{Median}\{|y_i - \hat{y}_i|\}_{i=1}^{N}$
    - robust, not sensitive to outliers
    - hard to optimise with gradients

Mean Absolute Error

$$\text{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

Characteristics

- Less sensitive to outliers
  - no squaring
- **Baseline:** median, not mean



Variants

- Median Absolute Error: $= \text{Median}\{|y_i - \hat{y}_i|\}_{i=1}^{N}$
  - robust, not sensitive to outliers
  - hard to optimise with gradients
- Median Squared Error: $= \text{Median}\{(y_i - \hat{y}_i)^2\}_{i=1}^{N}$

Correlation Coefficient

$$\mathrm{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \rho(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{\mathrm{Cov}(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\sqrt{\mathrm{Var}(\boldsymbol{y})}\sqrt{\mathrm{Var}(\hat{\boldsymbol{y}})}} \qquad \in [-1, 1]$$

Figures: Wikipedia

Correlation Coefficient

$$\mathrm{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \rho(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{\mathrm{Cov}(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\sqrt{\mathrm{Var}(\boldsymbol{y})}\sqrt{\mathrm{Var}(\hat{\boldsymbol{y}})}} \qquad \in [-1, 1]$$

Characteristics

- Insensitive to scaling and translation

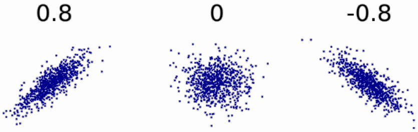
Figures: Wikipedia

Correlation Coefficient

$$\mathrm{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \rho(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{\mathrm{Cov}(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\sqrt{\mathrm{Var}(\boldsymbol{y})}\sqrt{\mathrm{Var}(\hat{\boldsymbol{y}})}} \qquad \in [-1, 1]$$

Characteristics

- Insensitive to scaling and translation
- No units



Showing correlation of y with respect to $\hat{y}$ for three different datasets

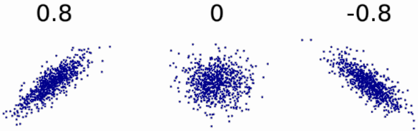
Figures: Wikipedia

Correlation Coefficient

$$\text{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \rho(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{\text{Cov}(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\sqrt{\text{Var}(\boldsymbol{y})}\sqrt{\text{Var}(\hat{\boldsymbol{y}})}} \qquad \in [-1, 1]$$

Characteristics

- Insensitive to scaling and translation
- No units
- Signals agreement on *relative ordering*



Showing correlation of y with respect to $\hat{y}$ for three different datasets

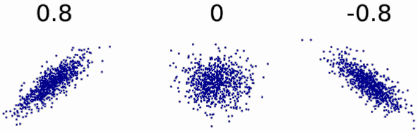
Figures: Wikipedia

Correlation Coefficient

$$\mathrm{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \rho(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{\mathrm{Cov}(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\sqrt{\mathrm{Var}(\boldsymbol{y})}\sqrt{\mathrm{Var}(\hat{\boldsymbol{y}})}} \qquad \in [-1, 1]$$

Characteristics

- Insensitive to scaling and translation
- No units
- Signals agreement on *relative ordering*
  - for larger y, predict larger $\hat{y}$



Showing correlation of y with respect to $\hat{y}$ for three different datasets

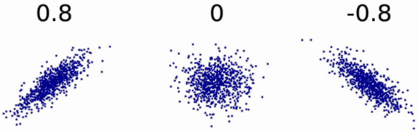
Figures: Wikipedia

Correlation Coefficient

$$\text{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \rho(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{\text{Cov}(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\sqrt{\text{Var}(\boldsymbol{y})}\sqrt{\text{Var}(\hat{\boldsymbol{y}})}} \qquad \in [-1, 1]$$

Characteristics

- Insensitive to scaling and translation
- No units
- Signals agreement on *relative ordering*
  - for larger y, predict larger $\hat{y}$
  - for smaller y, predict smaller $\hat{y}$



Showing correlation of y with respect to $\hat{y}$ for three different datasets

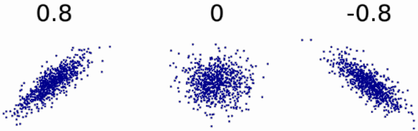
Figures: Wikipedia

Correlation Coefficient

$$\mathrm{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \rho(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{\mathrm{Cov}(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\sqrt{\mathrm{Var}(\boldsymbol{y})}\sqrt{\mathrm{Var}(\hat{\boldsymbol{y}})}} \qquad \in [-1, 1]$$

Characteristics

- Insensitive to scaling and translation
- No units
- Signals agreement on *relative ordering*
  - for larger y, predict larger $\hat{y}$
  - for smaller y, predict smaller $\hat{y}$
  - …or vice versa



Showing correlation of y with respect to $\hat{y}$ for three different datasets

Figures: Wikipedia

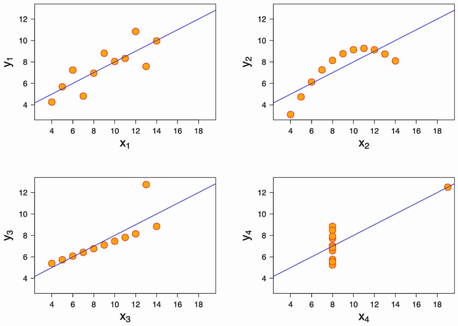
Correlation Coefficient

$$\operatorname{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \rho(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{\operatorname{Cov}(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\sqrt{\operatorname{Var}(\boldsymbol{y})}\sqrt{\operatorname{Var}(\hat{\boldsymbol{y}})}} \qquad \in [-1, 1]$$

Characteristics

- Insensitive to scaling and translation
- No units
- Signals agreement on *relative ordering*
  - for larger y, predict larger $\hat{y}$
  - for smaller y, predict smaller $\hat{y}$
  - …or vice versa
- Critical to visualise

  pitfalls: Anscombe's Quartet



Figures: Wikipedia

Coefficient of Determination (R^2)

$$\mathrm{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = 1 - \frac{\mathrm{MSE}(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\mathrm{Var}(\boldsymbol{y})} \qquad \in [-\infty, 1]$$

Coefficient of Determination (R^2)

$$\operatorname{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = 1 - \frac{\operatorname{MSE}(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\operatorname{Var}(\boldsymbol{y})} \qquad \in [-\infty, 1]$$

Characteristics

- R^2 measures goodness of fit for model

Coefficient of Determination (R^2)

$$\mathrm{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = 1 - \frac{\mathrm{MSE}(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\mathrm{Var}(\boldsymbol{y})} \qquad \in [-\infty, 1]$$

Characteristics

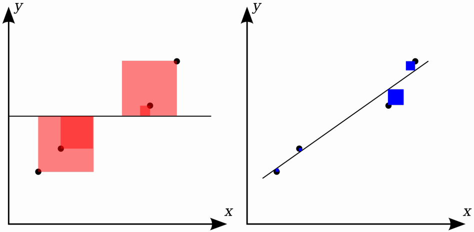
- R^2 measures goodness of fit for model
- How much variance in y is explained by the model?

Coefficient of Determination (R^2)

$$\text{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = 1 - \frac{\text{MSE}(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\text{Var}(\boldsymbol{y})} \qquad \in [-\infty, 1]$$

Characteristics

- R^2 measures goodness of fit for model
- How much variance in y is explained by the model?
- **Baseline:** predict the mean label!



$R^2 = 1$ (perfect fit)

$R^2 = 0$ (baseline predictor)

$R^2 < 0$ (incorrect model)

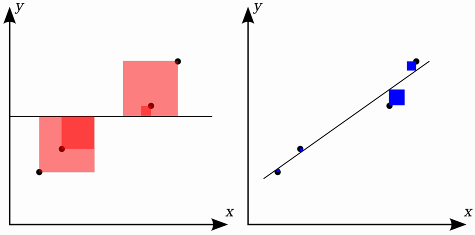
Coefficient of Determination (R^2)

$$\mathrm{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = 1 - \frac{\mathrm{MSE}(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\mathrm{Var}(\boldsymbol{y})} \qquad \in [-\infty, 1]$$

Characteristics

- R^2 measures goodness of fit for model
- How much variance in y is explained by the model?
- **Baseline:** predict the mean label!

$$\mathrm{Var}(\boldsymbol{y}) = \tfrac{1}{N} \textstyle\sum_{i=1}^{N} (y_i - \mu_y)^2$$



$R^2 = 1$ (perfect fit)

$R^2 = 0$ (baseline predictor)

$R^2 < 0$ (incorrect model)

Coefficient of Determination (R^2)

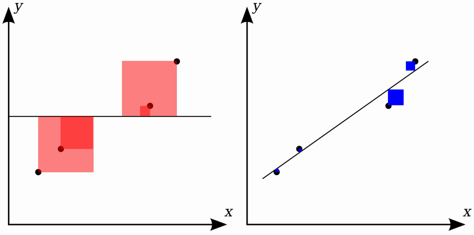
$$\mathrm{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = 1 - \frac{\mathrm{MSE}(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\mathrm{Var}(\boldsymbol{y})} \qquad \in [-\infty, 1]$$

Characteristics

- R^2 measures goodness of fit for model
- How much variance in y is explained by the model?
- **Baseline:** predict the mean label!

$$\mathrm{Var}(\boldsymbol{y}) = \tfrac{1}{N} \sum_{i=1}^{N} (y_i - \mu_y)^2$$

$$\mu_y = \tfrac{1}{N} \sum_{i=1}^{N} y_i$$



$R^2 = 1$ (perfect fit)

$R^2 = 0$ (baseline predictor)

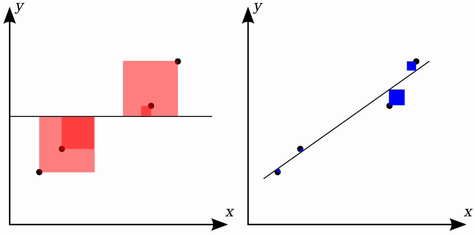
$R^2 < 0$ (incorrect model)

Coefficient of Determination (R^2)

$$\text{dist}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = 1 - \frac{\text{MSE}(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\text{Var}(\boldsymbol{y})} \qquad \in [-\infty, 1]$$

Characteristics

- R^2 measures goodness of fit for model
- How much variance in y is explained by the model?
- **Baseline:** predict the mean label!
  $\text{Var}(\boldsymbol{y}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \mu_y)^2$
  $\mu_y = \frac{1}{N} \sum_{i=1}^{N} y_i$
- Under some conditions $\rho^2 = R^2$!
  (convex optimality with scaling and translation)



$R^2 = 1$ (perfect fit)

$R^2 = 0$ (baseline predictor)

$R^2 < 0$ (incorrect model)

Summary

Evaluation

- Classification
    - Accuracy, TPR, FPR, Cohen’s Kappa
    - Precision-Recall, ROC curves

Summary

Evaluation

- Classification
    - Accuracy, TPR, FPR, Cohen’s Kappa
    - Precision-Recall, ROC curves

- Regression
    - MSE, MAE
    - Correlation, R^2