

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

Evaluation

Classification

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

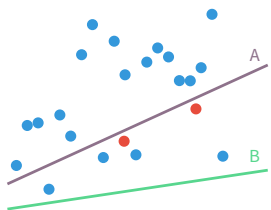
- $\text{Error} = \frac{\text{incorrect}}{\text{total}} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$
- $\text{Accuracy} = \frac{\text{correct}}{\text{total}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$

Measure of how “good” a classifier is

Issue: Class imbalances

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

Classification Accuracy: Class Imbalance



$\text{Acc}(\text{B}) > \text{Acc}(\text{A})!$

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

- False Positive Rate (FPR) = $\frac{\text{FP}}{\text{FP} + \text{TN}}$
 - (False Alarm) % of “—” misclassified as “+”
- False Negative Rate (FNR) = $\frac{\text{FN}}{\text{TP} + \text{FN}}$
 - (Miss) % of “+” misclassified as “—”
- Recall / True Positive Rate (TPR) = $\frac{\text{TP}}{\text{TP} + \text{FN}}$
 - (1 - Miss) % of “+” correctly predicted
- Precision / Positive Predictive Rate (PPR) = $\frac{\text{TP}}{\text{TP} + \text{FP}}$
 - % of “+” out of all positive predictions

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

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

Error Measures

Unified Measures:

optimisation objective, comparative evaluation

- Detection cost

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

- F-measure

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

- 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 $T = TP + FP + TN + FN$

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

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

Thresholds

- Models typically compute “confidence” as $p(y|x)$

- Decisions are made by **thresholding** this confidence

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

- Logistic regression: $\sigma(w^T x) \geq 0.5$

- Naive Bayes: $p(y = \text{spam}|x) > 0.5$

- τ determines error rates and confusion matrix
each τ provides a value for chosen measure(s)

- Complete picture:**

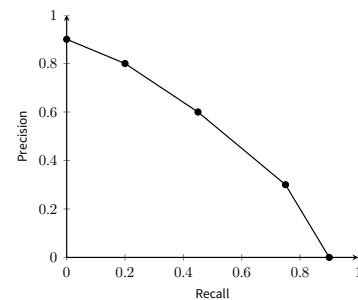
plot measures as τ varies from $-\infty$ to ∞

		Predicted	
		+	-
τ_1	True +	200	300
	True -	100	400
		Predicted	
		+	-
τ_2	True +	400	100
	True -	300	200
		...	

Precision-Recall Curve

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

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



email	label	$p(y x)$
“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

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

- Same positive recognition rate (66.7%)

- Very different negative recognition rates: **strong** on left, **nil** on right

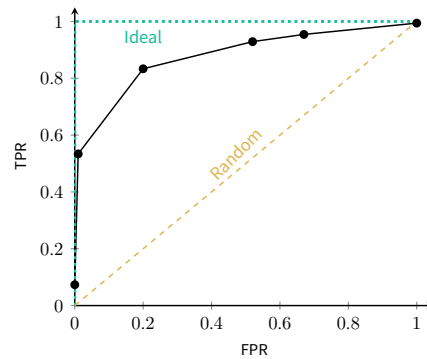
Receiver Operating Characteristic (ROC)

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

AUC

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

True \ Predicted	Predicted	
	+	-
+	TP 200	FN 300
-	FP 100	TN 400



Multi-Class Classification

Measures

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

$$MCC = \frac{p_o - p_e}{\sqrt{(1 - p_y)(1 - p_{\bar{y}})}}$$

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

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

True \ Predicted	Predicted							
	1	2	3	4	5	6	7	8
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

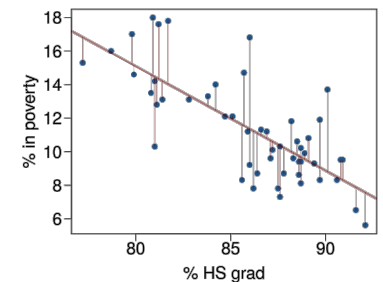
Evaluating Regression

Classification

- count how often incorrect

Regression

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



$$\text{dist}(\mathbf{y}, \hat{\mathbf{y}})$$

Evaluating Regression

$\text{dist}(\mathbf{y}, \hat{\mathbf{y}})$

Error Measures

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

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

Evaluating Regression

$\text{dist}(\mathbf{y}, \hat{\mathbf{y}})$

Error Measures

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

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

Evaluating Regression

$\text{dist}(\mathbf{y}, \hat{\mathbf{y}})$

Error Measures

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

$$\text{dist}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{\text{Cov}(\mathbf{y}, \hat{\mathbf{y}})}{\sqrt{\text{Var}(\mathbf{y})} \sqrt{\text{Var}(\hat{\mathbf{y}})}}$$

Evaluating Regression

$\text{dist}(\mathbf{y}, \hat{\mathbf{y}})$

Error Measures

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

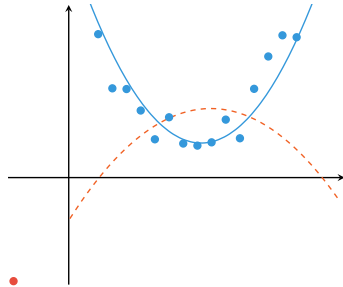
$$\text{dist}(\mathbf{y}, \hat{\mathbf{y}}) = 1 - \frac{\text{MSE}(\mathbf{y}, \hat{\mathbf{y}})}{\text{Var}(\mathbf{y})}$$

Mean Squared Error

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

Characteristics

- Sensitivity to outliers
 - squaring → blow up error

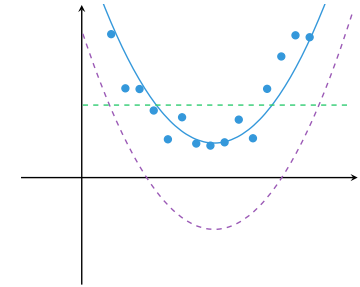


Mean Squared Error

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

Characteristics

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



Mean Absolute Error

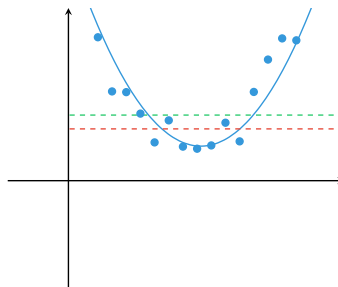
$$\text{dist}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

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: = $\text{Median}\{(y_i - \hat{y}_i)^2\}_{i=1}^N$

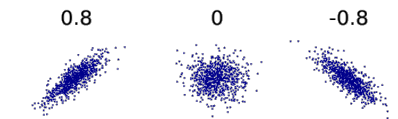


Correlation Coefficient

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

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



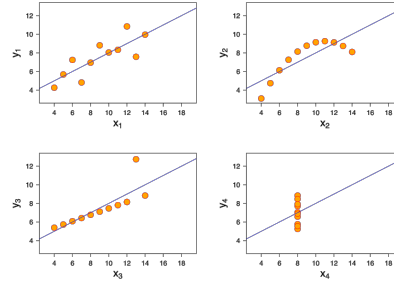
Showing correlation of y with respect to $\hat{y}$ for three different datasets

Correlation Coefficient

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

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

14

Coefficient of Determination (R^2)

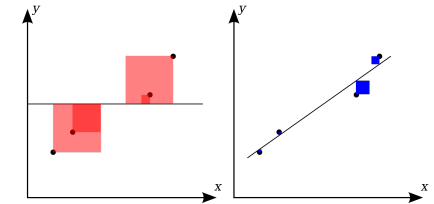
$$\text{dist}(\mathbf{y}, \hat{\mathbf{y}}) = 1 - \frac{\text{MSE}(\mathbf{y}, \hat{\mathbf{y}})}{\text{Var}(\mathbf{y})} \in [-\infty, 1]$$

Characteristics

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

$$\text{Var}(\mathbf{y}) = \frac{1}{N} \sum_{i=1}^N (y_i - \mu_y)^2$$

$$\mu_y = \frac{1}{N} \sum_{i=1}^N y_i$$
- 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)

15

Summary

Evaluation

- Classification
 - Accuracy, TPR, FPR, Cohen's Kappa
 - Precision-Recall, ROC curves
- Regression
 - MSE, MAE
 - Correlation, R^2

16