Computational Cognitive Science

Lecture 13: Active learning

Chris Lucas

School of Informatics

University of Edinburgh

29 October, 2024

Reading

Recommended:

- “Inferring causal networks from observations and interventions” by Steyvers et al. (2003)

Active learning

In our examples so far, including

- The direction-judgment task
- Categorization
- The number game
- Causal learning and attribution

We have assumed that people are passive observers.

Active learning

In reality, we tend to take an active role in gathering information:

- The direction-judgment task
  - looking at less-crowded, more-informative parts of the scene
- Categorization
  - choosing examples to get labels for
- The number game
  - asking if specific numbers are part of the concept
- Causal learning and attribution
  - intervening, designing experiments

Active learning

To count as active learning, there must be

1. selection of action or information that
2. serves some learning goal

Active learning

Some active learning is unconscious, e.g., gaze

Active learning

Other kinds are conscious, e.g.,

- Calling someone's bluff
- Squeezing a fruit
- Hefting an object
- Tuning a guitar
- Visiting a new restaurant
- Internet searches

Formalizing active learning

What does it mean to learn more or less; how can we quantify learning?

Active learning

Some intuitions:

- Learning reduces our uncertainty
- Learning changes our beliefs

Reducing uncertainty

How do we define uncertainty?

Suppose we want to know if someone has a particular illness.

- If we think it's $p = .5$ that they do, we are maximally uncertain.
- If we think it's $p = .99$ or 0.01, we're almost certain.
- If we think it's $p = 1$ or 0, we're certain.

Reducing uncertainty

Another perspective:

- If we are certain, and don't have much left to learn; we don't expect to be surprised.
- Let's define **surprisal** associated with event x as $\log(1/P(x)) = -\log(P(x))$

What if we formalize uncertainty as **expected surprisal**:

$$\mathbb{E}_{P(x)}[-\log(P(x))]$$

Reducing ~~uncertainty~~ entropy

Expected surprisal is **entropy**, the standard way to quantify uncertainty.

$$H(X) = -\sum_{x \in \mathcal{X}} P(x) \log P(x)$$

(log base 2 $\rightarrow$ bits; $e \rightarrow$ "nats")

Reducing ~~uncertainty~~ entropy

Uncertain:

$$p = .5 \rightarrow \; -(0.5 * \log .5 + 0.5 \log .5) = 1 \text{ bit}$$

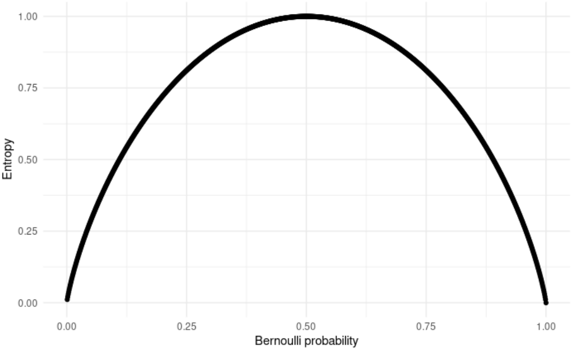
Almost certain:

$$p = .99 \rightarrow \; -(.99 * \log .99 + .01 \log .01) = 0.08 \text{ bit}$$

Certain:

$$p = 1 \rightarrow \; -(1 * \log 1 + 0 \log 0) = 0 \text{ bit}$$

Entropy



1.00
0.75
0.50
0.25
0.00
Entropy
0.00
0.25
0.50
0.75
1.00
Bernoulli probability

Entropy

We can think of entropy as the amount of information we expect to need to be certain about a variable's value.

Entropy and communication: Example

Alice randomly chooses a candy from an bag with the following mix:

- 1/2 Anise candies
- 1/8 Blackcurrant candies
- 1/8 Chocolate candies
- 1/8 Dulce de leche candies
- 1/8 Earl grey candies

Alice wants to tell us what candy she has by blinking (left=0, right=1).

How many bits (blinks) does she need?

Entropy

Claude Shannon showed that you can expect to need at least

$$-\sum_{x\in\mathcal{X}} P(x)\log P(x)$$

bits of information (the entropy).

As noted earlier, any event x has a **surprisal** (or "information content") $I(x) = -\log P(x)$

Entropy

What are our surprisals for Alice?

$$-\log(1/2, 1/8, 1/8, 1/8, 1/8) \to 1, 3, 3, 3, 3$$

Expected surprisal:

$$4 * (1/8) * 3 + 1 * (1/2) * 1 = 1.5 + .5 = 2 \text{ bits}$$

Entropy

How does this translate to actual communication?

For our candies, Alice could use the following code:

- A ($p = 1/2$) $\rightarrow$ 0
- B ($p = 1/8$) $\rightarrow$ 111
- C ($p = 1/8$) $\rightarrow$ 110
- D ($p = 1/8$) $\rightarrow$ 101
- E ($p = 1/8$) $\rightarrow$ 100

Half the time she'll need 1 bit. Half the time she'll need 3 bits.

Entropy

That is,

$$0.5 * 1 + 0.5 * 3 = 2 \text{ bits}.$$

on average.

That's the entropy of the candy-choice random variable – we can't do better.

Entropy and active learning

We want to choose an action a that reduces our expected uncertainty, i.e., our entropy.

This is can expressed as *information gain*:

$$H(P(y)) - \mathbb{E}_{P(d|a)}[H(P(y|a,d))]$$

- y is what we care about
- a is our action
- d is the unknown result of our action

We're communicating with the universe, but

- we can't agree a code in advance and
- we don't know how informative its message will be.

Entropy and active learning

How do we maximize our information gain?

$$H(P(y)) - \mathbb{E}_{P(d|a)}[H(P(y|a,d))]$$

We want to minimize

$$\mathbb{E}_{P(d|a)}[H(P(y|a,d))] = \sum_{d \in \mathcal{D}} P(d|a) H(y|a,d)$$

That is, we want to pick actions that are probably going to be informative.

Mutual information and KL divergence

We can also represent MI as:

- How much our data are expected to change our posterior beliefs relative to our priors (as measured by KL divergence).
- How much the joint distribution $p(X, Y)$ differs from the joint distribution assuming X and Y are independent.

Focusing on uncertain events

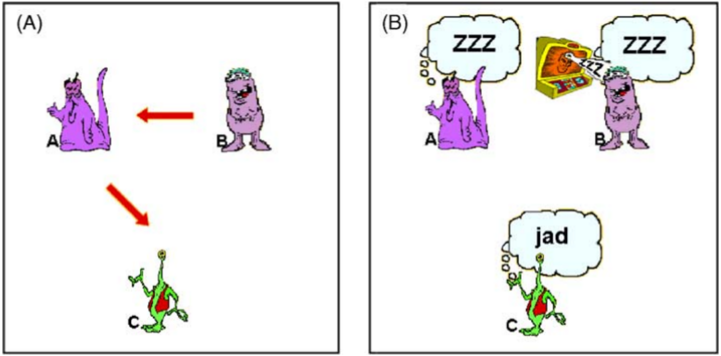
We learn little from experiments where we know what the outcome will be.

What is we focus on experiments where we don't know what will happen?

- "maximum entropy sampling" – sometimes very useful, but if misleading if some observations are inherently noisy
- sometimes a pitfall for new scientists!

Example: Alien mind reading

Steyvers et al. (2003) asked whether people choose causal interventions to in learn in an efficient way, using an information-gain approach.



Example: Alien mind reading

Participants saw 18 kinds of acyclic causal graphs and made causal structure judgments based on

- Observations
- Interventions with a brain zapper

The causal relationships were stochastic – aliens could fail to read other minds.

Example: Alien mind reading

Steyvers tested different active learning models by manipulating the hypothesis space:

- **Rational identification**: $\mathcal{H}$ includes all possible hypotheses
- **Rational test 1**: A working hypothesis versus a null hypothesis (independence)
- **Rational test 2**: A working hypothesis versus a simpler model with one fewer edge

Overall:

- The **rational test** models fit people better than identification
- *Some* participants might have been using strategies resembling rational identification

Some additional observations

Steyvers et al. considered only 3 variables and 18 possible structures.

It seems implausible that people do anything like rational identification when there are many variables.

Some additional observations

We have been talking about optimizing the information gain from a **single** observation – a **greedy** or **myopic** policy.

In general, many observations are necessary for learning, and myopic policies are rarely optimal overall.

Non-myopic optimal policies tend to be so expensive that cognitive scientists don't bother with them and call myopic policies optimal.

Myopia

Suppose we want to distinguish between:

- U: a 60/40 bias coin, with numbers on opposite sides that have even sums
- V: a 55/45 bias coin, with numbers that have odd sums

We have three action options:

- Flip the coin
- Look at the head-face serial number
- Look at the tail-face serial number

If we only look at the informativeness of individual actions, we will flip the coins **many** times.

If we can look at the total informativeness of sets of 2+ actions, our entropy will be zero after we check both faces.

Example: Wason's card selection task

We want to know the truth of the rule "If there is a vowel on one side of a card, there is an even number on the opposite side", given

- One card showing E
- One card showing C
- One card showing 8
- One card Showing 3

What card should we turn over?

Wason: This is a logic puzzle that people failed by choosing 8.

Oaksford and Chater (1994) treated this as an active learning problem

Example: Wason's card selection task

How do we gain information about the truth of the rule?

Oaksford and Chater turned to tripe-eating as a more intuitive cover story:

Rule: "If you eat tripe, you will feel ill."

Four alternatives:

- P: Ask a person who ate tripe if they feel ill: Both outcomes are informative
- !P: Ask a tripe-avoider if they feel ill: Useless
- Q: Ask an ill person if they ate tripe: potentially useful, depending on how common tripe-avoidance and illness are
- !Q: Ask a well person if they ate tripe: potentially useful, depending on how common tripe-avoidance and illness are

Example: Wason's card selection task

If tripe-eating is common, !Q gives us a chance to decisively answer the question If tripe-eating is rare, we are likely to learn nothing

If illness is rare, Q could help us substantiate the rule If illness is rare, Q is less helpful

Example: Wason's card selection task

Oakford and Chater showed that human behavior is consistent with "optimal data selection"

aka active learning.

Other models

We have focused on a specific (myopically) rational model. Other models exist, based on various heuristics, e.g.,

- Positive test strategies, which tend to search for evidence consistent with a hypothesis, at the expense of falsification
- Divide-and-conquer strategies, which try to eliminate 50% (or nearly) of hypotheses
- Predictive-coding based approaches, which often resemble maximum entropy sampling

References

Murphy, K. (2022). Probabilistic Machine Learning: An introduction. MIT Press. (Yes, 2022; see https://probml.github.io/pml-book/book1.html). Section 19.4.

Oaksford, M. & Chater, N. (1994). "A rational analysis of the selection task as optimal data selection". Psychological Review. 101 (4): 608–631

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. Cognitive Science, 27(3), 453–489