Information criteria

ISTA 410 / INFO 510: Bayesian Modeling and Inference

U. of Arizona School of Information

October 12, 2020

Outline

Last week:

- Posterior predictive checking
- Prior predictive checking
- Graphical assessment of model performance

This week:

- Information theory and predictive accuracy
- Scoring models to avoid overfitting

A few ideas from information theory

Uncertainty

The main contribution of information theory to statistics is a measurable notion of uncertainty.

What is uncertainty?

- We don't know the value of future observations yet

Key notion: surprise

- We are surprised by observing rare events
- Surprising events carry high information

Uncertainty: example

A simple example of varying uncertainty: weather forecasts.

- Today's weather in Tucson: sunny.
  Tomorrow's weather?

Uncertainty: example

A simple example of varying uncertainty: weather forecasts.

- Today's weather in Tucson: sunny.
  Tomorrow's weather?
- Today's weather in Seattle: cloudy.
  Tomorrow's weather?

Uncertainty: example

A simple example of varying uncertainty: weather forecasts.

- Today's weather in Tucson: sunny.
  Tomorrow's weather?
- Today's weather in Seattle: cloudy.
  Tomorrow's weather?
- Today's weather in Chicago: rainy.
  Tomorrow's weather?

Information entropy

Measurement for uncertainty: *information entropy.* Introduced by Claude Shannon (1947) at Bell Labs.

p any probability distribution:

$$H(p) = -\sum_i p_i \log_2(p_i)$$

Measurement of uncertainty is average negative log probability. (The negative log probability is the "surprise" or Shannon information of each event.)

Key property: maximized by flat distributions.

Entropy and encoding

Basic application of entropy: symbol codes

- Goal: encode information (e.g. text messages) into sequences of bits (0/1)
- Assign a bit string (called a code word) to each symbol in the alphabet
- How many bits does each symbol need?

Entropy and encoding

Basic application of entropy: symbol codes

- Goal: encode information (e.g. text messages) into sequences of bits (0/1)
- Assign a bit string (called a code word) to each symbol in the alphabet
- How many bits does each symbol need?
- Exploit symbol frequencies: assign shorter code words to more common symbols
- Theoretical minimum *average* length: entropy of the frequency distribution

Kullback-Leibler divergence

Kullback-Leibler (KL) divergence:

$$D_{KL}(p, q) = \sum p_i(\log_2 p_i - \log_2 q_i)$$

Interpretation:

- p is the true outcome distribution
- q is the model predictive distribution
- KL divergence measures incorrectness, in some way

Potential for surprise

A really useful interpretation of KL divergence is as a “potential for surprise.” (The idea of surprise as a measurable quantity is all over information theory.)

Imagine two scenarios:

- You raise a dog in Chicago, and then you move here to Tucson
- You raise a dog in Tucson, and then you move to Chicago

Potential for surprise

The weather in Chicago is variable:

- It's hot and humid in the summer
- It's bitterly cold in the winter
- Sometimes it just oscillates between the two on a daily basis

Your Chicagoan dog has experienced all kinds of weather, and will be comfortable in the heat and the cold

Potential for surprise

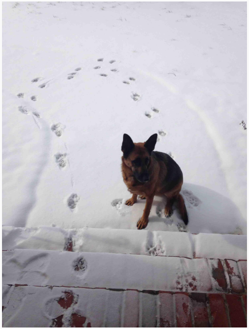
As previously noted, the weather in Tucson is pretty consistent.

Your Tucsonan dog, upon moving to Chicago:

Potential for surprise

As previously noted, the weather in Tucson is pretty consistent.

Your Tucsonan dog, upon moving to Chicago:



Asymmetry in KL divergence

This is reflected by the asymmetry in KL divergence.

City	Tucson	Chicago
p_{hot}	0.95	0.5
p_{cold}	0.05	0.5

$$D_{KL}(\mathrm{Tuc}, \mathrm{Chi}) = 0.714$$

$$D_{KL}(\mathrm{Chi}, \mathrm{Tuc}) = 1.198$$

Asymmetry in KL divergence

Statistical interpretation:

- A flat model is closer to a nonflat model than vice versa
- Advantage to simpler models: they have higher entropy

Cross entropy

KL divergence is closely related to another measurement, *cross-entropy*:

$$H(p, q) = -\sum_i p_i \log_2(q_i)$$

Immediately:

$$D_{KL}(q, p) = H(q, p) - H(p)$$

Application and interpretation of cross-entropy

Cross-entropy has a nice interpretation in the encoding context: if you construct an optimal symbol code for a frequency distribution p, and you use it to encode text coming from an alphabet with a frequency distribution q, the cross entropy is the expected length.

Common application: target function for ML classifiers

Information criteria for scoring models

Log score and deviance

Scoring models using log probabilities:

$$\text{log score} \qquad S(q) = \sum \log(q_i)$$

Deviance:

$$D = -2 * S(q) = -2 * \sum \log(q_i)$$

In Bayesian world, the posterior isn't one model, it's a distribution of models – so we should average:

$$\text{lppd}(y, \theta) = \sum_i \log \frac{1}{S} \sum_s p(y_i|\theta_s)$$

Out-of-sample prediction error

The problem is to estimate prediction error (evaluated by log score) out of sample.

- Adding parameters generally improves fit within the sample
- Eventually, adding parameters reduces accuracy out of the sample (overfitting)
- How can we predict out-of-sample prediction accuracy?
    - Cross-validation
    - Information criteria

Akaike information criterion (AIC)

AIC: named for Akaike (but he called it "an information criterion")

Assuming a point estimate $\hat{\theta}$ for model parameters, calculate the log score and apply a penalty to correct for overfitting:

$$AIC = -2 \log p(y|\hat{\theta}_{\text{mle}}) + 2k$$

k is the number of parameters. Assumes Gaussian posterior.

Where it comes from: Taylor expansion around the posterior mode.

Widely applicable information criterion

Introduced by Watanabe (2010); a more Bayesian generalization of the AIC.

$$WAIC = -2(\mathrm{lppd}(y, \theta) - \sum_i \mathrm{var} \log(p(y_i|\theta)))$$

lppd replaces the training deviance; pointwise variance in log posterior density generalizes the parameter count (effective number of parameters).

Reduces to AIC in the special case where AIC is exact: models with flat priors and Gaussian posterior.

Summary

Today:

- information theory / entropy
- information criteria: AIC / WAIC

- Applied examples
- Cross-validation methods