Computational Cognitive Science

Lecture 7: Model comparison and selection

Chris Lucas

School of Informatics

University of Edinburgh

October 8, 2019

Readings

- Chapter 10 of F&L
- “Ockham’s razor and Bayesian Analysis” (link)

Recommended:

- “A note on the evidence and Bayesian Occam’s razor” (link)

Model comparison

We have discussed estimating parameters conditional on a model.

- That may be all we need, if we can capture different theories as parameter choices in a single model
- In practice, we may want to compare qualitatively different models

How do we choose between models?

Criteria for choosing models

We prefer models that are

1. Predictively useful
2. Compatible with our data
3. Likely to be correct, or closer to a correct model

(Understandable, too)

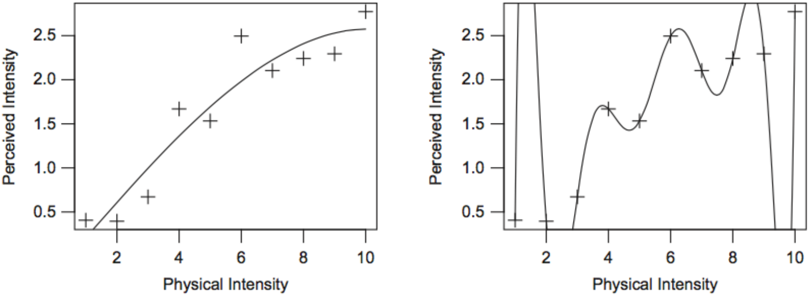
Two models of perceived intensity

- $\mathcal{M}_1$: Perceived intensity is a **2nd** order polynomial function of physical intensity
- $\mathcal{M}_2$: Perceived intensity is a **9th** order polynomial function of physical intensity

(Ignore the fact that we could distinguish between these models w/a single model and suitable priors over parameters)

Two models of perceived intensity

Both models, with MLE fits[1]:



Which is better?

[1] Figure 10.1 in F&L.

Two models

- Is the complex polynomial going to give good predictions?
  - $p(y_{K+1}|\mathbf{y}, \mathcal{M}_2)$
- Is the complex polynomial compatible with our data?
  - $p(\mathbf{y}|\mathcal{M}_2)$
- Is the complex polynomial the right generative model??
  - $p(\mathcal{M}_2|\mathbf{y})$

An important distinction:

- A **specific** 9th order polynomial, versus
- **some** 9th order polynomial.

Predictive accuracy

- Is the complex polynomial going to give good predictions?
  - $p(y_{K+1}|\mathbf{y}, \mathcal{M}_2)$

Suppose we have a model where all we care about is RMSE, and we can only obtain point-estimate predictions.

Are there any principles that should guide how we define a model?

Geman et al.[2] described *bias-variance dilemma*, explaining why "tabula rasa" models are not desirable.

[2] "Neural networks and the bias-variance dilemma" (1992) by Geman, Bienentock, and Doursat.

Bias and variance

- The expected RMSE of a regression model can be decomposed:
    - Error due to *bias*: The difference between the expected predictions of the model (under all possible data) and the real mean
    - Error due to *variance*: How much the model's predictions vary as a function of the specific data it has been given

Bias and variance

- The ideal model:
  - predictions are matched to reality (in expectation); no bias-based error
  - predictions don't depend on idiosyncrasies of data; no variance-based error
  - Extreme version: A perfectly confident and accurate prior
- Highly flexible models will do poorly unless large data sets are available

The lesson: If we have a priori information or constraints, we should use them!

Two models

For probababilistic models, predictive accuracy relates to other desiderata:

- Is the complex polynomial compatible with our data?
  - $p(\mathbf{y}|\mathcal{M}_2)$
- Is the complex polynomial the right generative model?
  - $p(\mathcal{M}_2|\mathbf{y}) \propto p(\mathbf{y}|\mathcal{M}_2)P(\mathcal{M}_2)$

To answer these questions, we need the *marginal likelihood* of our data.

Two models

Marginal likelihood:

$$\begin{aligned} p(\mathbf{y}|\mathcal{M}) &= \int_{\boldsymbol{\theta}} p(\mathbf{y}|\mathcal{M},\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{M})d\boldsymbol{\theta} \\ &\neq p(\mathbf{y}|\mathcal{M},\boldsymbol{\theta}) \end{aligned}$$

- Flexible models can accommodate a wide variety of patterns
- If those patterns are not present in our data, they're bad models

Flexibility and overfitting: Likelihood

What if we specify $p(\boldsymbol{\theta})$ at the start, and compute $p(\mathbf{y}|\mathcal{M})$?

That's an excellent solution, when it's viable.

However:

- We must choose priors carefully
- Integrating over $\boldsymbol{\theta}$ is often expensive or impossible

Model comparison without marginal likelihood

What if we can't compute the marginal likelihood, but can compute likelihoods and MLEs?

- Compare predictive accuracy/likelihood on held-out test data

Model comparison without marginal likelihood

What if we don't have a test set?

- E.g., using a data set where alternative models were fitted to the whole set
- Very few data points, s.t., estimating parameters already difficult

Three common approaches:

1. Likelihood ratios vs χ^2
2. AIC and BIC
3. Cross-validation

Nested models and χ^2

Suppose $\mathcal{M}_1$ is a special case of $\mathcal{M}_2$; $\mathcal{M}_2$ has additional parameters and reduces to $\mathcal{M}_1$ for specific values of these parameters. We can say $\mathcal{M}_1$ is *nested* in $\mathcal{M}_2$.

Even if the additional parameters of $\mathcal{M}_2$ are useless – they just allow it to fit noise – the negative log likelihood will be slightly lower.

Nested models and likelihood ratios

However, under certain assumptions and as *n* goes to infinity, that improvement (times 2) will converge to a χ^2 distribution with df equal to the difference in dimensionality[3].

As a result, one can compare the difference in MLE likelihoods to a χ^2 distribution to support reject the hypothesis that the complex model is no better.

$$2 \cdot [\log(p(\mathbf{y}|\hat{\boldsymbol{\theta}}_2, \mathcal{M}_2) - \log(p(\mathbf{y}|\hat{\boldsymbol{\theta}}_1, \mathcal{M}_1)]$$

Caveats:

- If models are nested, there are often nice Bayesian approaches
- Null hypothesis significance test

[3] To learn more, see Wilks' theorem (link)

AIC

Another approach: “How different is the distribution implied by my model from the real-world distribution of human behavior?”

How can we quantify this difference?

Kullback-Leibler divergence[4]:

$$\int_{\mathbf{y}} R(\mathbf{y}) \log \frac{R(\mathbf{y})}{p_M(\mathbf{y})} d\mathbf{y}$$

.

If these distributions are identical, divergence is zero. If the model assigns zero probability density to events that are possible, it’s ∞.

[4] Wikipedia article. Don’t call it a distance.

AIC

AIC approximates relative KL divergences of models to target distribution (e.g., relative probabilities of behaviors):

$$\mathrm{AIC} = 2k - 2 \cdot \log(p(\mathbf{y}|\boldsymbol{\theta}_{MLE}))$$

- Asymptotically agrees with leave-one-out cross-validation
- There are many alternatives, but AIC is simple and popular

AIC

Caveats:

- Approximates hold-one-out cross-validation, not extrapolation
- Approximation is asymptotic; not necessarily great for small data sets
- Parameter counting is sometime a poor way to evaluate complexity; see text
- Cross-validation makes fewer assumptions, is intuitive and robust – generally better
- Consider alternatives like AIC_C

Prediction (again)

The best way to assess a model's predictive accuracy: Predict with it

Prediction (again)

The best way to assess a model's predictive accuracy: Predict with it

1. Sequester a subset of your data. Don't touch it. Don't look at it. Pretend it doesn't exist
    - To see if a model can predict the judgments or behavior of new participants or in new conditions, hold out participants and/or conditions
    - Likewise for future judgments given past judgments

Prediction (again)

The best way to assess a model's predictive accuracy: Predict with it

2. Fit models on separate data, compare their predicitive log likelihoods on the sequestered data
    - No need to penalize model complexity

Cross-validation

If you want robust and efficient estimates of predictive accuracy, you can repeat those steps for your entire data set;

- Don't look at *anything* before building the model
- Define an automatic policy for partitioning and fitting the model
- Repeat for K "folds" (train on $K-1$, evaluate on 1)
- Offers approximate predictive likelihoods for new folds

In practice, cognitive scientists rarely use fully held-out test sets.

- Tend to look at data when tuning model
- Cross-validation with seen-data is still better than testing and training on the same data

Summary

If we want to choose between models, we can do the following:

1. Compare marginal likelihoods
   - Easy in concept, difficult (sometimes impossible) in practice
2. Compare predictive loss with fully held-out evaluation set(s)
   - In practice, typically just one partitioning
3. Compare predictive losses w/cross-validation
   - A pragmatic approach given sparse data
   - Mitigates the worst of the "train on test" problem
   - Good partitionings require care
4. AIC or likelihood-ratio test
   - Blunt instrument, but common
   - See also the AIC_C, BIC, WAIC, ...