Model Comparison
Machine Learning and Pattern Recognition

Chris Williams

School of Informatics, University of Edinburgh

October 2015

(These slides have been adapted from previous versions by Charles Sutton, Amos Storkey and David Barber)
Overview

- The model selection problem
- Overfitting
- Validation set, cross validation
- Bayesian Model Comparison
- Reading: Murphy 1.4.7, 1.4.8, 6.5.3, 5.3; Barber 12.1-12.4, 13.2 up to end of 13.2.2
Model Selection

- We may entertain different models for a dataset, $M_1$, $M_2$, \ldots, e.g. different numbers of basis functions, different regularization parameters
- How should we choose amongst them?
- Example from supervised learning

**Linear Regression**

**Cubic Regression**

**9th-Order Regression**
Loss and Training Error

- For input $x$ the true target is $y(x)$ and our prediction is $f(x)$. The loss function
  \[ L(y(x), f(x)) \]
  assesses errors in prediction.

- Examples
  - squared error loss $(y(x) - f(x))^2$,
  - 0-1 loss $I(y(x), f(x))$ for classification,
  - log loss $-\log p(y(x)|f(x))$ (probabilistic predictions)

- Training error
  \[ E_{tr} = \frac{1}{N} \sum_{n=1}^{N} L(y(x^n), f(x^n)) \]

- Training error consistently decreases with model complexity
Overfitting

- Generalization (or test) error

\[ E_{gen} = \int L(y(x), f(x)) \, p(x, y) \, dx \, dy \]

- Overfitting (Mitchell 1997, p. 67)

  A hypothesis \( f \) is said to **overfit** the data if there exists some alternative hypothesis \( f' \) such that \( f \) has a smaller training error than \( f' \), but \( f' \) has a smaller generalization error than \( f \).
Partition the available data into two: a training set (for fitting the model), and a validation set (aka hold-out set) for assessing performance.

Estimate the generalization error with

\[ E_{val} = \frac{1}{V} \sum_{v=1}^{V} L(y(x^v), f(x^v)) \]

where we sum over cases in the validation set.

Unbiased estimator of the generalization error

Suggested split: 70% training, 30% validation
Cross Validation

➤ Split the data into $K$ pieces (folds)
➤ Train on $K - 1$, test on the remaining fold
➤ Cycle through, using each fold for testing once
➤ Uses all data for testing, cf. the hold-out method

Figure credit: Murphy Fig 1.21(b)
Cross Validation: Example

- Degree 14 polynomial with $N = 21$ datapoints
- Regularization term $\lambda w^T w$
- How to choose $\lambda$?

Figure credit: Murphy Fig 7.7
- Left-hand end of $x$-axis $\equiv$ low regularization
- Notice that training error increases monotonically with $\lambda$
- Minimum of test error is for an intermediate value of $\lambda$
- Both cross validation and a Bayesian procedure (coming soon) choose regularized models
Bayesian Model Comparison

▸ Have a set of different possible models

\[ M_i \equiv p(D|\theta, M_i) \text{ and } p(\theta|M_i) \]

for \( i = 1, \ldots, K \)

▸ Each model is set of distributions that have associated parameters. Usually some models are more complex (have more parameters) than others

▸ Bayesian way: Have a prior \( p(M_i) \) over the set of models \( M_i \), then compute posterior \( p(M_i|D) \) using Bayes’ rule

\[
p(M_i|D) = \frac{p(M_i)p(D|M_i)}{\sum_{j=1}^{K} p(M_j)p(D|M_j)}
\]

▸

\[ p(D|M) = \int p(D|\theta, M)p(\theta|M) \, d\theta \]

This is called the *marginal likelihood* or the *evidence*. 
Comparing models

Bayes factor = \( \frac{P(D|M_1)}{P(D|M_2)} \)

\[
\frac{P(M_1|D)}{P(M_2|D)} = \frac{P(M_1)}{P(M_2)} \cdot \frac{P(D|M_1)}{P(D|M_2)}
\]

Posterior ratio = Prior ratio \( \times \) Bayes factor

Strength of evidence from Bayes factor (Kass, 1995; after Jeffreys, 1961)

1 to 3 \hspace{1cm} \text{Not worth more than a bare mention}
3 to 20 \hspace{1cm} \text{Positive}
20 to 150 \hspace{1cm} \text{Strong}
> 150 \hspace{1cm} \text{Very strong}
Computing the Marginal Likelihood

- Exact for conjugate exponential models, e.g. beta-binomial, Dirichlet-multinomial, Gaussian-Gaussian (for fixed variances)

- E.g. for Dirichlet-multinomial

\[
p(D|M) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + N) \prod_{i=1}^{r} \frac{\Gamma(\alpha_i + N_i)}{\Gamma(\alpha_i)}}
\]

- Also exact for (generalized) linear regression (for fixed prior and noise variances)

- Otherwise various approximations (analytic and Monte Carlo) are possible
BIC approximation

\[ \text{BIC} = \log p(D|\hat{\theta}) - \frac{\text{dof}(\hat{\theta})}{2} \log N \]

- Bayesian information criterion (Schwarz, 1978)
- \( \hat{\theta} \) is MLE
- \( \text{dof}(\hat{\theta}) \) is the degrees of freedom in the model (\( \sim \) number of parameters in the model)
- BIC penalizes ML score by a penalty term
- BIC is quite a crude approximation to the marginal likelihood
Why Bayesian model selection? Why not compute best fit parameters and compare?

More parameters = better fit to data. ML: bigger is better.

But might be overfitting: only these parameters work. Many others don’t.

Prefer models that are unlikely to ‘accidentally’ explain the data.
Example

You are an auditor of a firm. You receive details about the sales that a particular salesman is making. He attempts to make 4 sales a day to independent companies. You receive a list of the number of sales by this agent made on a number of days. Explain why you would expect the total number of sales to be binomially distributed.

If the agent was making the sales numbers up as part of a fraud, you might expect the agent (as he is a bit dim) to choose the number of sales at random from a uniform distribution. You are aware of the fraud possibility, and you understand there is something like a $1/5$ chance this salesman is involved.

Given daily sales counts of $1 2 2 4 1 4 3 2 4 1 3 3 2 4 3 3 2 3 3$, do you think the salesman is lying?
Binomial Example

Example

Data: 1 2 2 4 1 4 3 2 4 1 3 3 2 4 3 3 2 3 3

- $\mathcal{M} = 1$ - From $P_1(x|p)$ a binomial distribution $\text{Binomial}(4)$. Prior on $p$ is uniform.
- $\mathcal{M} = 2$ - From $P_2(x)$ a uniform distribution $\text{Uniform}(0, \ldots, 4)$.
- Discuss what you would do?
- $P(\mathcal{M} = 1) = 0.8$. 
Example

Data: 1 2 2 4 1 4 3 2 4 1 3 3 2 4 3 3 2 3 3

- $\mathcal{M} = 1$ - From $P_1(x|p)$ a binomial distribution $\text{Binomial}(4)$. Prior on $p$ is uniform.
- $\mathcal{M} = 2$ - From $P_2(x)$ a uniform distribution $\text{Uniform}(0, \ldots, 4)$.
- $P(\mathcal{M} = 1) = 0.8$.

$$P(D|\mathcal{M} = 1) = \int dp \, P_1(D|p)P(p) \quad , \quad P(D|\mathcal{M} = 2) = P_2(D)$$

$$P(\mathcal{M}|D) = \frac{P(D|\mathcal{M})P(\mathcal{M})}{P(D|\mathcal{M} = 1)P(\mathcal{M} = 1) + P(D|\mathcal{M} = 2)P(\mathcal{M} = 2)}$$

- Left as an exercise! (see tutorial)
Linear Regression Example

- $d=1$, $\log ev=-18.593$, EB
- $d=2$, $\log ev=-20.218$, EB
- $d=3$, $\log ev=-21.718$, EB

$N=5$, method=EB

$P(M|D)$
Summary

- Training and test error, overfitting
- Validation set, cross validation
- Bayesian Model Comparison