Model Comparison

Machine Learning and Pattern Recognition

Chris Williams

School of Informatics, University of Edinburgh

October 2014

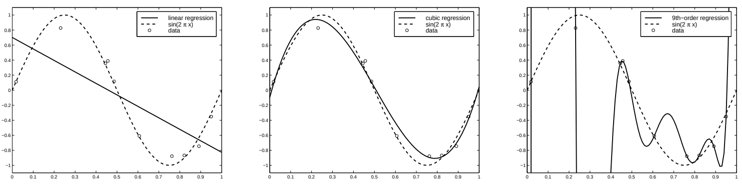
(These slides have been adapted from previous versions by Charles Sutton, Amos Storkey and David Barber

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, ..., e.g. different numbers of basis functions, different regularization parameters
- How should we choose amongst them?
- Example from supervised learning



linear cubic 9th-order

Loss and Training Error

- For input $\mathbf{x}$ the true target is $y(\mathbf{x})$ and our prediction is $f(\mathbf{x})$. The loss function

$$L(y(\mathbf{x}), f(\mathbf{x}))$$

assesses errors in prediction
- Examples
  - squared error loss $(y(\mathbf{x}) - f(\mathbf{x}))^2$,
  - 0-1 loss $I(y(\mathbf{x}), f(\mathbf{x}))$ for classification,
  - log loss $-\log p(y(\mathbf{x})|f(\mathbf{x}))$ (probabilistic predictions)
- Training error

$$E_{tr} = \frac{1}{N} \sum_{n=1}^{N} L(y(\mathbf{x}^n), f(\mathbf{x}^n))$$

- Training error consistently decreases with model complexity

Overfitting

- Generalization (or test) error

$$E_{gen} = \int L(y(\mathbf{x}), f(\mathbf{x}))\; p(\mathbf{x}, y)\; d\mathbf{x}\; dy$$

- 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.

Validation Set

- 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

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

  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

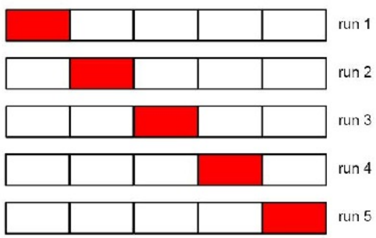


Figure credit: Murphy Fig 1.21(b)

Cross Validation: Example

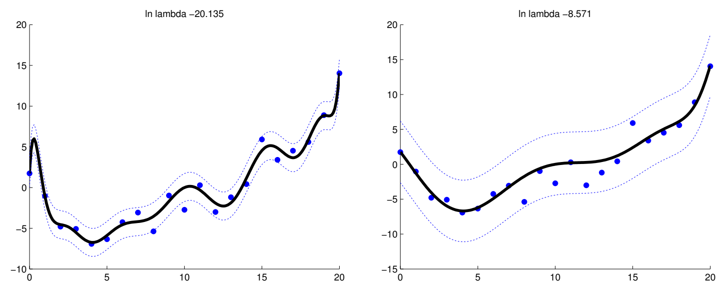


Figure credit: Murphy Fig 7.7

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

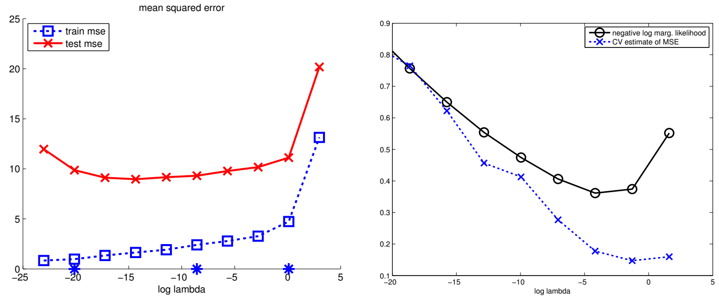


Figure credit: Murphy Fig 7.7

- Left-hand end of x-axis $\equiv$ low regularization
- Notice that training error increases monotonically with λ
- Miminum of test error is for an intermediate value of λ
- Both cross validation and a Bayesian procudure (coming soon) choose regularized models

Bayesian Model Comparison

- Have a set of different possible models

$$\mathcal{M}_i \equiv p(\mathcal{D}|\theta, M_i) \text{ and } p(\theta|M_i)$$

  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 $p(M_i|\mathcal{D})$ using Bayes' rule

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

-

$$p(\mathcal{D}|M) = \int p(\mathcal{D}|\theta, M)p(\theta|M)\, d\theta$$

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

Comparing models

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

$$\frac{P(M_1|\mathcal{D})}{P(M_2|\mathcal{D})} = \frac{P(M_1)}{P(M_2)} . \frac{P(\mathcal{D}|M_1)}{P(\mathcal{D}|M_2)}$$

$$\text{Posterior ratio} = \text{Prior ratio} \times \text{Bayes factor}$$

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

1 to 3	Not worth more than a bare mention
3 to 20	Positive
20 to 150	Strong
> 150	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

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

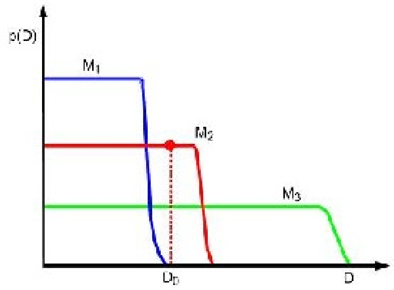
- Also exact for (generalized) linear regression (for fixed prior and noise variances)
- Otherwise various approximations (analytic and Monte Carlo) are possible

BIC approximation

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

- Bayesian information criterion (Schwarz, 1978)
- $\hat{\theta}$ is MLE
- 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.

Binomial Example

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

Binomial Example

Example

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

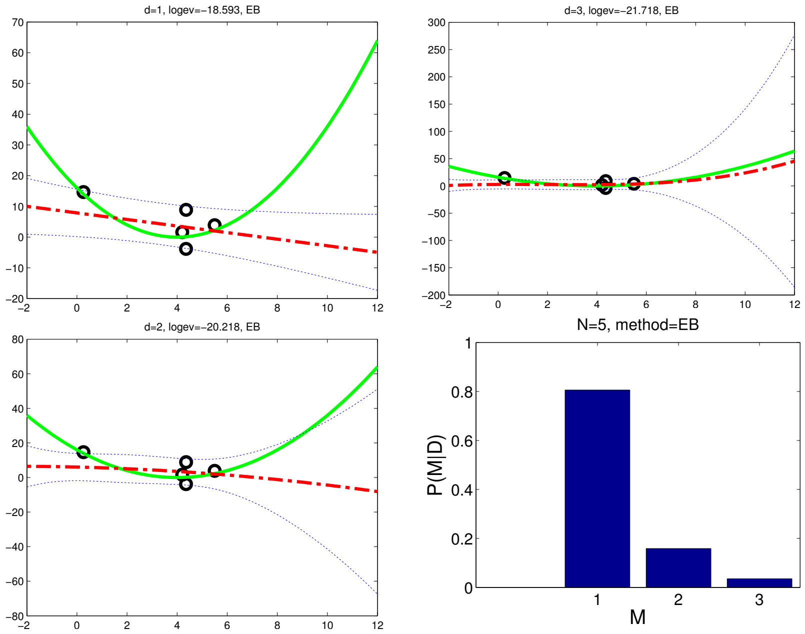
- $\mathcal{M} = 1$ - From $P_1(x|p)$ a binomial distribution Binomial(4). Prior on p is uniform.
- $\mathcal{M} = 2$ - From $P_2(x)$ a uniform distribution Uniform(0,. . . ,4).
- $P(\mathcal{M} = 1) = 0.8$.

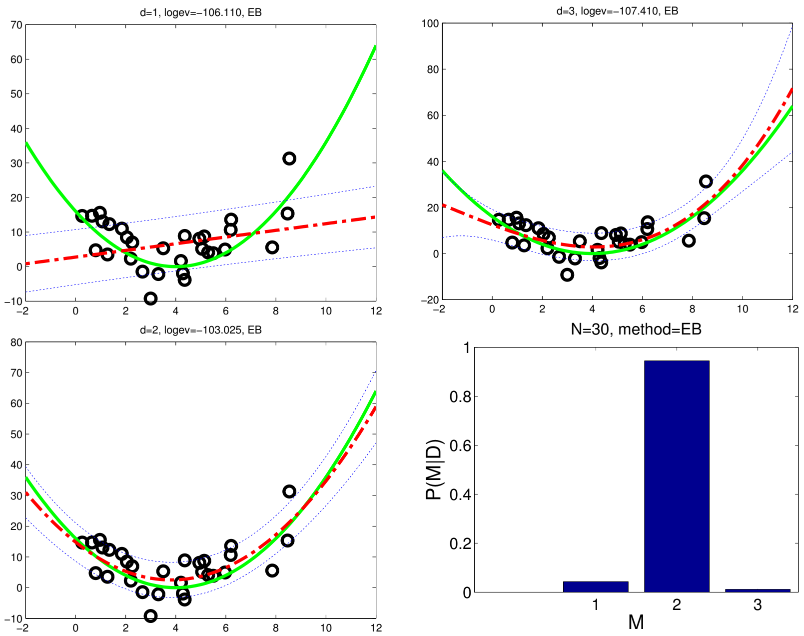
$$P(\mathcal{D}|\mathcal{M} = 1) = \int dp\ P_1(\mathcal{D}|p)P(p)\ ,\ P(\mathcal{D}|\mathcal{M} = 2) = P_2(\mathcal{D})$$

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

- Left as an exercise! (see tutorial)

Linear Regression Example





Summary

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