Model Selection

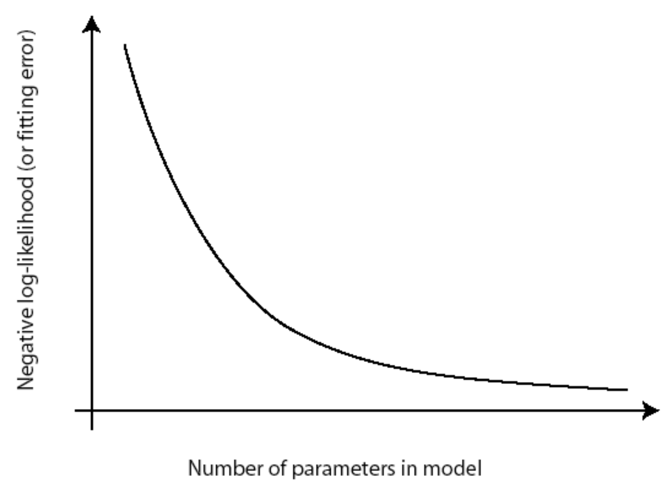
- Model selection refers to choosing among different instances within a model class (1) or different model classes (2).
- Examples:
  - The number of clusters (1)
  - The degree of a polynomial to fit a curve to data (1)
  - Polynomials versus other basis functions such as Fourier (2)

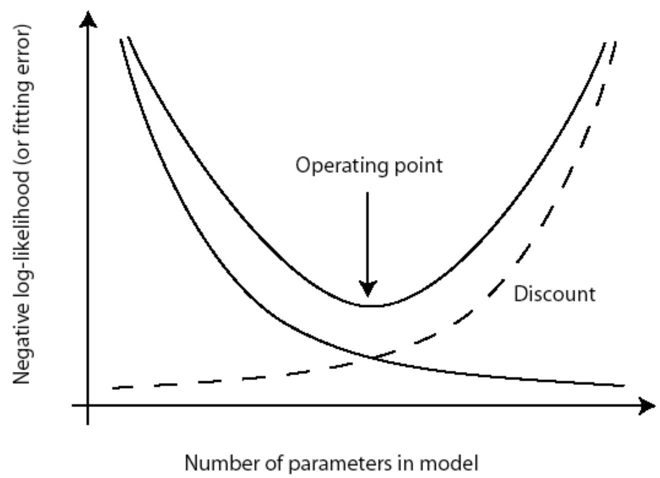
Model Evaluation

- There is no real difference between choosing the parameters of a model (fitting), which model class instance (learning), and which model class (structure learning).
- Difficulties
  - Prior densities of different models are typically of different dimensionality (leads to expensive integration).
  - Likelihoods can often help select models, but constructing them is an exacting task.
  - Good priors over different model classes are often not obvious

Solutions (penalize complexity)

- Typical approach is to focus on the balance between fitting accuracy, and model complexity using various penalties.





Solutions (penalize complexity)

- Typical approach is to focus on the balance between fitting accuracy, and model complexity using various penalties.
- AIC (An information criterion, Akaike, 74)

Replace log likelihood, $\log(\mathrm{p}(\mathrm{D}|\theta))$, with $\log(\mathrm{p}(\mathrm{D}|\theta)) - M$ where M is the number of adjustable parameters.

Solutions (penalize complexity)

- Typical approach is to focus on the balance between fitting accuracy, and model complexity using various penalties.
- BIC (Bayesian information criterion)

Replace log likelihood, $\log(\mathrm{p}(\mathrm{D}|\theta))$, with $\log(\mathrm{p}(\mathrm{D}|\theta)) - \frac{1}{2} M \log(N)$ where M is the number of adjustable parameters, N is the number of data points. This is the usual approximation. See Bishop, page 216-217 for a more complicated version.

Often also called minimum discription length (MDL)

Solutions (likelihood function)

- Incorrect complex models may predict lots of data where there is none
- Solution is to model missing data
- Example --- finding asteroids from detections amidst noise

Solutions (likelihood function)

- Incorrect complex models may predict lots of data where there is none
- Solution is to model missing data
- Example --- finding asteroids from detections amidst noise
  - Predicting more asteroids explains more data, but we expect to see detections where detected
  - Good modeling the probability of noise detections and probability of missing detections has a greater affect on the posterior than a prior (necessarily not very strong) on the number of asteroids.

Solutions (integrating parameter uncertainty)

$p(D|M_i) = \int p(D|\theta)p(\theta|M_i)d\theta$ (Model evidence)

and we can evaluate $p(M_i|D)$ by Bayes.

Note that the dimension of the space of θ is typically a function of i.

This is argued (Bishop, §3.4) to be a principled way to penalize complex models because complex models spread their probability mass over greater support.

Under additional approximations and assumptions, this becomes BIC (Bishop, §4.4.1).

Solutions (model averaging)

Recall the predictive distributions

$$p(x \mid X) = \int p(x \mid \theta)p(\theta \mid X)d\theta$$

To mitigate uncertainty of different models

$$p(x|X) = \sum_i p(M_i) \int p(x \mid \theta_i)p(\theta_i \mid X, M_i)d\theta$$

Note typical assumption that M_i influences x through θ_i only, so no conditioning on M_i in the first factor in the integral.

Cross-validation

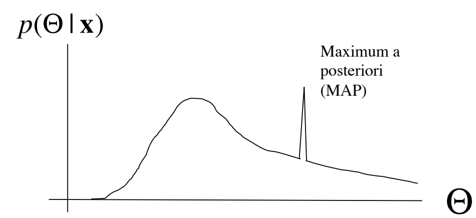
- Standard way to evaluate models
- Exclude a subset of the data while fitting model
- Compute predictions for the held-out subset.
- Evaluate predictions against actual held-out values
  - e.g., distance from truth, or class labels
- If you use k such sets, this is called k-fold cross-validation
- If you leave out 1 data point, it is called leave-one-out.

Cross-validation (2)

- Cross-validation provides
  - A way to choose models
  - A way to measure performance
  - A way to measure generalization capacity
- Held out data **must be different enough** to test the level of generality that you want
  - Consider degree of validation in a model to predict happiness
    1. How happy are you now given recent data points
    2. How happy are you now given all data points
    3. How happy are you on day X given data for other days
    4. How happy are you based on model of **other** people
    5. How happy are you based on **other** people in other experiments
    6. How happy are you based on modeling people in other cultures

More on estimation

- If the goal is to provide the model, then we typically estimate the MAP value for the parameters
- This assumes that the posterior is nicely behaved
- An alternative is to average some or all (MMSE) of the posterior.



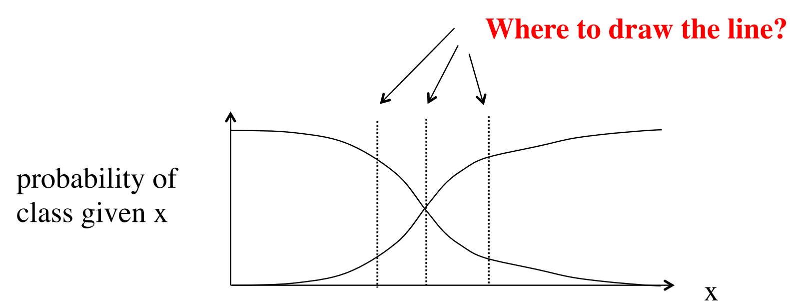
Classification

- Consider that our parameters include a discrete class variable, c.
- Assume no other variables, or that they have been marginalized out.
- Use x for the data.

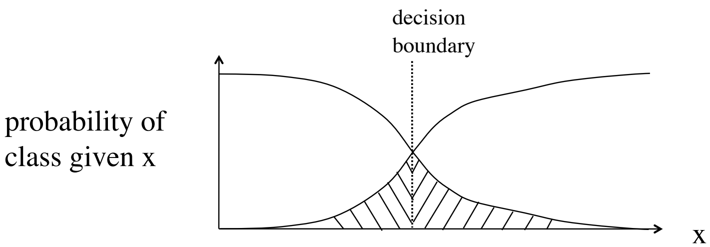
$$p(c \mid x) \propto p(c)p(x \mid c)$$

- So, given x, what is the class?

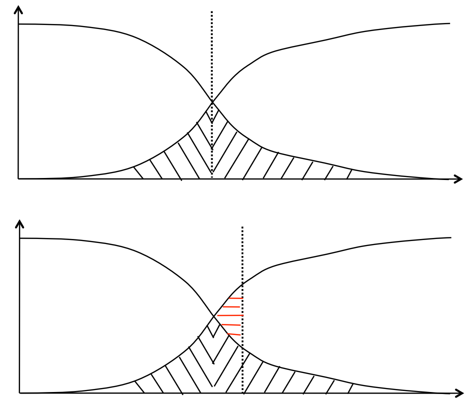
Classification



Basic ideas in classification



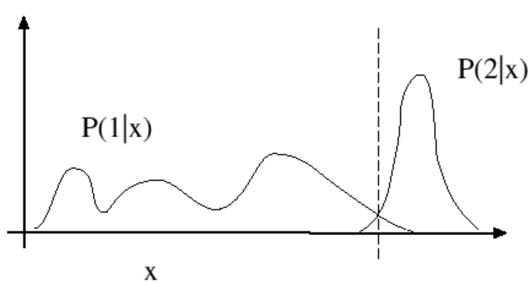
Area of intersection under curves gives expected value of making a mistake



Red shows extra that you get wrong with different boundary

Classification

Finding a decision boundary is not the same as modeling a conditional density.



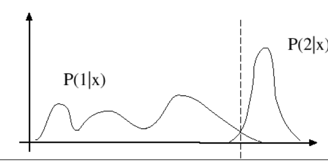
Classification

Finding a decision boundary is not the same as modeling a conditional density.

Working with the boundary might be easier (we don't care about the extra bumps).

But we loose any indication of whether the point is an outlier.

In this course we will not consider the approach of finding the boundary (discriminative method).



Decision making

Classification where the risk (loss) for each class is different.

Example: Risk of a false negative diagnosis is more than that for the risk of false positive diagnosis.

Define a loss function, Lj,k which tells us the loss of classifying a category k, as a category, j.

Example:

$$\begin{array}{cc} & \begin{array}{cc}\text{cancer} & \text{normal}\end{array} \\ \begin{array}{c}\text{cancer} \\ \text{normal}\end{array} & \begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix} \end{array}$$

Decision making

Now set classification boundaries for **x** to minimize expected loss.

This is done by minimizing:

$$\sum_k L_{k,j} \cdot p\left(C_k | x\right)$$

Decision making

Example to illustrate that the formula is sensible.

Suppose that at a given x*, we have

$p\left(C_1 \mid x^*\right) = 0.3 \quad p\left(C_2 \mid x^*\right) = 0.2 \quad p\left(C_3 \mid x^*\right) = 0.5$

Evaluate the assignment of x* under loss functions

$$\mathrm{L}_A = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \qquad \mathrm{L}_B = \begin{pmatrix} 0 & 1 & 1 \\ 10 & 0 & 10 \\ 1 & 1 & 0 \end{pmatrix}$$

$p\left(C_1 \mid x^*\right) = 0.3 \quad p\left(C_2 \mid x^*\right) = 0.2 \quad p\left(C_3 \mid x^*\right) = 0.5$

For the first example (loss is misclassifcation rate)

$$\mathrm{L}_B = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

Note that loss is defined for misclassifying the column item as the row item.

Declaring that x at x* is C_1 has loss: (0.3)*0 +(0.2)*1+ (0.5)*1=0.7
Declaring that x at x* is C_2 has loss: (0.3)*1 +(0.2)*0+ (0.5)*1=0.8
Declaring that x at x* is C_3 has loss: (0.3)*1 +(0.2)*1+ (0.5)*0=0.5

As expected, the minimum loss is for the likeliest class.

$p(C_1 | x^*) = 0.3 \quad p(C_2 | x^*) = 0.2 \quad p(C_3 | x^*) = 0.5$

For the second example

$$L_B = \begin{pmatrix} 0 & 1 & 1 \\ 10 & 0 & 10 \\ 1 & 1 & 0 \end{pmatrix}$$

Note that loss is defined for misclassifying the column item as the row item.

Declaring that x at x* is C_1 has loss: (0.3)*0 +(0.2)*10+ (0.5)*1 = 2.5

Declaring that x at x* is C_2 has loss: (0.3)*1 +(0.2)*0+ (0.5)*1 = 0.8

Declaring that x at x* is C_3 has loss: (0.3)*1 +(0.2)*10+ (0.5)*0 = 2.3

Now the heavy loss of missing C_2 leads to C_2 being the best answer.
(Note that C_2 was the worst answer with the previous loss).