Cross-validation

• Split data set into two pieces, fit to one, and compute negative log-likelihood on the other
• One set is “training data”, the other is “testing data” or “held out data”
• Average over different splits
• This estimates the quality of your model
  – Often (rightfully so) used to compare algorithms
• If you are doing model selection, then you choose the model with the smallest value of this average
  – This works because adding parameters causes over fitting of the training data which gives worse performance on test data

Model averaging

• Often smarter to use multiple models for prediction than just one
• Consider that we have various models that we believe to various degrees, denoted by \( P(M_i) \)
• Suppose we want to estimate \( X \) from data, \( D \), via the group of models, \( M_i \)
• A Bayesian would compute

\[
P(X | D) = \sum_i P(X | M_i, D) P(M_i | D)
\]

Recognition by finding patterns

• Template matching with correlation (linear filters) is a simple example of recognition by pattern matching
• Some objects behave like quite simple templates
  – Frontal faces

Recognition by finding patterns

• Example strategy:
  – Find image windows
  – Correct for lighting
  – Pass them to a statistical test (a classifier) that accepts faces and rejects non-faces
• Important high level point:
  – Need to understand relationship between modeling statistics and deciding between options (classification AND risk analysis).
Basic ideas in classification

- Concrete example
  - "guess" male / female from height

- Probabilistic approach
  - Consider $P(\text{female} | \text{height})$

Where to draw the line?

Area of intersection under curves gives expected value of making a mistake
Basic ideas in classification

- Concrete example
  - "guess" male / female from height
- Probabilistic approach
  - Consider P(female|height)
- Now consider "risk"
  - Suppose you want to give vaccine based on height for a disease that only males get.
  - There is great benefit to males who may be exposed
  - Vaccines have risk as well as benefit
  - Thus there is also some risk to giving females a vaccine they do not need
- How does this change the boundary?

Loss / Risk

- Some errors may be more expensive than others
  - e.g. a fatal disease that is easily cured by a cheap medicine with minimal side-effects → false positives in diagnosis are better than false negatives
- We want to set the classification point
- Consider two class classification
  - Let L(1→2) be the loss caused by calling a 1 a 2
  - Want to analyze the expected value of the loss (risk)

Basic ideas in classification

- Expected loss (risk) of using classifier $s$

$$R(s) = \Pr (1 \rightarrow 2 \mid \text{using } s)L(1 \rightarrow 2) + \Pr (2 \rightarrow 1 \mid \text{using } s)L(2 \rightarrow 1)$$

Details of formula optional, but the idea is worth understanding
Basic ideas in classification

• Generally, we should classify as 1 if the expected loss of classifying as 1 is better than for 2
• We get

1 if \( \Pr (1 \mid x)L(1 \rightarrow 2) > \Pr (2 \mid x)L(2 \rightarrow 1) \)
2 if \( \Pr (2 \mid x)L(2 \rightarrow 1) > \Pr (1 \mid x)L(1 \rightarrow 2) \)

• Crucial notion: Decision boundary
  – points where the loss is the same for either case

Details of formula optional, but the idea is worth understanding

Basic ideas in classification

• Expected loss (risk) of using classifier \( s \)

\[ R(s) = L(1 \rightarrow 2) \cdot \Pr(1 \rightarrow 2)(s) + L(2 \rightarrow 1) \cdot \Pr(2 \rightarrow 1)(s) \]

Skipped in 2008

Basic ideas in classification

• Expected loss (risk) of using classifier \( s \)

\[ R(s) = L(1 \rightarrow 2) \cdot \Pr(1 \rightarrow 2)(s) + L(2 \rightarrow 1) \cdot \Pr(2 \rightarrow 1)(s) \]

• So, given the class conditional densities, how do we set the boundary?

Where is the boundary?

• Expected loss (risk) of using classifier \( s \)

\[ R(s) = L(1 \rightarrow 2) \cdot \Pr(1 \rightarrow 2)(s) + L(2 \rightarrow 1) \cdot \Pr(2 \rightarrow 1)(s) \]

• Suppose that \( s \) is now our decision boundary, so \( x < s \leftrightarrow 1; x > s \leftrightarrow 2 \)
• So, what is \( \Pr(1 \rightarrow 2)(s) \)?
Where is the boundary?

• Expected loss (risk) of using classifier $s$
  \[ R(s) = L(1 \rightarrow 2) \cdot P(1 \rightarrow 2)(s) + L(2 \rightarrow 1) \cdot P(2 \rightarrow 1)(s) \]
  
• Suppose that $s$ is now our decision boundary, so $x < s \implies 1$; $x > s \implies 2$

• So, what is $P(1 \rightarrow 2)(s)$?
  \[ P(1 \rightarrow 2)(s) = \int_{s}^{\infty} p(1, x) \, dx \]

Small $p()$ reminds us that this is a probability density function, not a true probability. We use either $P()$ or $Pr()$ to emphasize that we have a true probability. Further confusion arises because probability and probability density are very close if the domain is discrete (e.g. two classes).
Where is the boundary?

- Expected loss (risk) is then
  \[ R(s) = \int_{s} L(1 > 2)p(1, x)dx + \int_{s} L(2 > 1)p(2, x)dx \]

- We want to minimize this. So differentiate and set to 0.
  \[
  \frac{d}{ds} \int_{s} L(1 > 2)p(1, x)dx = -L(1 > 2)p(1, s)
  \]
  \[
  \frac{d}{ds} \int_{s} L(2 > 1)p(2, x)dx = L(2 > 1)p(2, s)
  \]

(follows from the definition of integration and differentiation)

Basic ideas in classification

- Put differently
  \[
  1 \text{ if } P(1|x)L(1 \rightarrow 2) > P(2|x)L(2 \rightarrow 1) \\
  2 \text{ if } P(2|x)L(2 \rightarrow 1) > P(1|x)L(1 \rightarrow 2)
  \]

- (Switching to conditional probability is OK here)

- Crucial intuitive notion: Decision boundary is at the points where the loss is the same for either case