Decision Theory
Naïve Bayes
ROC Curves
Generative vs Discriminative Methods

• Logistic regression: $h: x \rightarrow y$.
• When we only learn a mapping $x \rightarrow y$ it is called a discriminative method.

• Generative methods learn $p(x,y) = p(x|y) p(y)$, i.e. for every class we learn a model over the input distribution.

• Advantage: leads to regularization for small datasets (but when $N$ is large discriminative methods tend to work better).
• We can easily combine various sources of information: say we have learned a model for attribute I, and now receive additional information about attribute II, then:
  $$p(x_I, x_{II} | y) \approx p(x_I | y) p(x_{II} | y)$$
• Disadvantage: you model more than necessary for making decisions, and input space ($x$-space) can be very high dimensional.

• This is called “conditional independence of $x|y$”.
• The corresponding classifier is called “Naïve Bayes Classifier”.

Naïve Bayes: decisions

\[ p(y \mid x_1, x_{II}) = \frac{p(x_1 \mid y)p(x_{II} \mid y)p(y)}{p(x_1, x_{II})} = \frac{p(x_1 \mid y)p(x_{II} \mid y)p(y)}{\sum_y p(x_1 \mid y)p(x_{II} \mid y)p(y)} \]

- This is the “posterior distribution” and it can be used to make a decision on what label to assign to a new data-case.

- Note that to make a decision you do not need the denominator.

- If we computed the posterior \( p(y \mid x_i) \) first, we can use it as a new prior for the new information \( x_{II} \) (prove this as home):

\[ p(y \mid x_1, x_{II}) \propto p(x_{II} \mid y)p(y \mid x_1) \]
Naïve Bayes: learning

• What do we need to learn from data?
  • \( p(y) \)
  • \( p(x_k|y) \) for all \( k \)

• A very simple rule is to look at the frequencies in the data: (assuming discrete states)
  \[ p(y) = \frac{\text{nr. data-cases with label } y}{\text{total nr. data-cases}} \]
  \[ p(x_k=i|y) = \frac{\text{nr. data-cases in state } x_k=i \text{ and } y}{\text{nr. data-cases with label } y} \]

• To regularize we imagine that each state \( i \) has a small fractional number of data-cases to begin with (\( K = \text{total nr. of classes} \)).
  \[ p(x_k=i|y) = \frac{c + \text{nr. data-cases in state } x_k=i \text{ and } y}{Kc + \text{nr. data-cases with label } y} \]

• What difficulties do you expect if we do not assume conditional independence?
• Does NB over-estimate or under-estimate the uncertainty of its predictions?

• Practical guideline: work in log-domain:
  \[ \prod_j p(x_j | y) \rightarrow \sum_j \log p(x_j | y) \]
Loss functions

- What if it is much more costly to make an error on predicting y=1 vs y=0?
- Example: y=1 is “patient has cancer”, y=0 means “patient is healthy.
- Introduce “expected loss function”:

\[ E[L] = \sum_{kj} L_{kj} \int_{R_j} dx \ p(y=k,x) \]

<table>
<thead>
<tr>
<th>Predict</th>
<th>cancer</th>
<th>healthy</th>
</tr>
</thead>
<tbody>
<tr>
<td>cancer</td>
<td>0</td>
<td>1000</td>
</tr>
<tr>
<td>healthy</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Total probability of predicting class while true class is k.

R_j is the region of x-space where an example is assigned to class j.
Decision surface

\[ E[L] = \sum_{kj} L_{kj} \int_{R_j} dx \ p(y = k, x) \]

- How shall we choose \( R_j \) ?

- Solution: minimize \( E[L] \) over \( \{R_j\} \).
  
  - Take an arbitrary point “x”.
  
  - Compute \( \sum_k L_{kj} p(y = k | x) \) for all \( j \) and maximize over “j”.

  - Since we maximize for every “x” separately, the total integral is maximal

  - Places where the decision switches belong to the “decision surface”.

- What matrix \( L \) corresponds to the decision rule on slide 2 using the posterior?
ROC Curve

- Assume 2 classes and 1 attribute.
- Plot class conditional densities \( p(x_k | y) \)
- Shift decision boundary from right to left.
- As you move the loss will change, so you want to find the point where it is minimized.

- If \( L = [0 \ 1; 1 \ 0] \) where is \( L \) minimal?
- As you shift the true true positive rate (TP) and the false positive rate (FP) change.
- By plotting the entire curve you can see the tradeoffs.
- Easily generalized to more attributes if you can find a decision threshold to vary.
Evaluation: ROC curves

TP = true positive rate = # positives classified as positive divided by # positives

FP = false positive rate = # negatives classified as positives divided by # negatives

TN = true negative rate = # negatives classified as negatives divided by # negatives

FN = false negatives = # positives classified as negative divided by # positives

Identify a threshold in your classifier that you can shift.

Plot ROC curve while you shift that parameter.