Logistic Regression

Required reading:
• Mitchell draft chapter (see course website)

Recommended reading:
• Bishop, Chapter 3.1.3, 3.1.4
• Ng and Jordan paper (see course website)
Naïve Bayes: What you should know

• Designing classifiers based on Bayes rule

• Conditional independence
  – What it is
  – Why it’s important

• Naïve Bayes assumption and its consequences
  – Which (and how many) parameters must be estimated under different generative models (different forms for \( P(X|Y) \) )

• How to train Naïve Bayes classifiers
  – MLE and MAP estimates
  – with discrete and/or continuous inputs
Generative vs. Discriminative Classifiers

Wish to learn $f: X \rightarrow Y$, or $P(Y|X)$

Generative classifiers (e.g., Naïve Bayes):

- Assume some functional form for $P(X|Y)$, $P(Y)$
  - This is the ‘generative’ model
- Estimate parameters of $P(X|Y)$, $P(Y)$ directly from training data
- Use Bayes rule to calculate $P(Y|X= x_i)$

Discriminative classifiers:

- Assume some functional form for $P(Y|X)$
  - This is the ‘discriminative’ model
- Estimate parameters of $P(Y|X)$ directly from training data
• Consider learning \( f: X \rightarrow Y \), where
  • \( X \) is a vector of real-valued features, \(< X_1 \ldots X_n >\)
  • \( Y \) is boolean
• We could use a Gaussian Naïve Bayes classifier
  • assume all \( X_i \) are conditionally independent given \( Y \)
  • model \( P(X_i | Y = y_k) \) as Gaussian \( N(\mu_{ik}, \sigma) \)
  • model \( P(Y) \) as Bernoulli \( (\pi) \)
• What does that imply about the form of \( P(Y|X) \)?
• Consider learning \( f: X \rightarrow Y \), where
  
  • \( X \) is a vector of real-valued features, \( < X_1 \ldots X_n > \)
  
  • \( Y \) is boolean
  
  • assume all \( X_i \) are conditionally independent given \( Y \)
  
  • model \( P(X_i | Y = y_k) \) as Gaussian \( N(\mu_{ik}, \sigma_i) \)
  
  • model \( P(Y) \) as Bernoulli \( (\pi) \)

• What does that imply about the form of \( P(Y|X) \)?

\[
P(Y = 1 | X = <X_1, \ldots X_n>) = \frac{1}{1 + exp(w_0 + \sum_i w_i X_i)}
\]
Very convenient!

\[ P(Y = 1|X = \langle X_1, \ldots, X_n \rangle) = \frac{1}{1 + \exp(w_0 + \sum_i w_i X_i)} \]

implies

\[ P(Y = 0|X = \langle X_1, \ldots, X_n \rangle) = \frac{\exp(w_0 + \sum_i w_i X_i)}{1 + \exp(w_0 + \sum_i w_i X_i)} \]

implies

\[ \frac{P(Y = 0|X)}{P(Y = 1|X)} = \exp(w_0 + \sum_i w_i X_i) \]

implies

\[ \ln \frac{P(Y = 0|X)}{P(Y = 1|X)} = w_0 + \sum_i w_i X_i \]

linear classification rule!
Derive form for $P(Y|X)$ for continuous $X_i$

\[
P(Y = 1|X) = \frac{P(Y = 1)P(X|Y = 1)}{P(Y = 1)P(X|Y = 1) + P(Y = 0)P(X|Y = 0)}
\]

\[
= \frac{1}{1 + \frac{P(Y = 0)P(X|Y = 0)}{P(Y = 1)P(X|Y = 1)}}
\]

\[
= \frac{1}{1 + \exp(\ln\frac{P(Y = 0)P(X|Y = 0)}{P(Y = 1)P(X|Y = 1)})}
\]

\[
= \frac{1}{1 + \exp\left(\ln\frac{\mu_{i0}}{\pi} + \sum_i \ln\frac{P(X_i|Y = 0)}{P(X_i|Y = 1)}\right)}
\]

\[
P(Y = 1|X) = \frac{1}{1 + \exp\left(w_0 + \sum_{i=1}^{n} w_i X_i\right)}
\]
Very convenient!

\[ P(Y = 1|X = \langle X_1, ... X_n \rangle) = \frac{1}{1 + \exp(w_0 + \sum_i w_i X_i)} \]

implies

\[ P(Y = 0|X = \langle X_1, ... X_n \rangle) = \frac{\exp(w_0 + \sum_i w_i X_i)}{1 + \exp(w_0 + \sum_i w_i X_i)} \]

implies

\[ \frac{P(Y = 0|X)}{P(Y = 1|X)} = \exp(w_0 + \sum_i w_i X_i) \]

implies

\[ \ln \frac{P(Y = 0|X)}{P(Y = 1|X)} = w_0 + \sum_i w_i X_i \]
Logistic function

\[ P(Y = 1|X) = \frac{1}{1 + \exp(w_0 + \sum_{i=1}^{n} w_i X_i)} \]
Logistic regression more generally

- Logistic regression in more general case, where \( Y \in \{Y_1 \ldots Y_R\} \): learn \( R-1 \) sets of weights

for \( k < R \)

\[
P(Y = y_k|X) = \frac{\exp(w_{k0} + \sum_{i=1}^{n} w_{ki}X_i)}{1 + \sum_{j=1}^{R-1} \exp(w_{j0} + \sum_{i=1}^{n} w_{ji}X_i)}
\]

for \( k = R \)

\[
P(Y = y_R|X) = \frac{1}{1 + \sum_{j=1}^{R-1} \exp(w_{j0} + \sum_{i=1}^{n} w_{ji}X_i)}
\]
Training Logistic Regression: MCLE

• Choose parameters $W = \langle w_0, \ldots w_n \rangle$ to maximize conditional likelihood of training data

$$P(Y = 0 | X, W) = \frac{1}{1 + \exp(w_0 + \sum_i w_i x_i)}$$

$$P(Y = 1 | X, W) = \frac{\exp(w_0 + \sum_i w_i x_i)}{1 + \exp(w_0 + \sum_i w_i x_i)}$$

• Training data $D = \{ \langle X^1, Y^1 \rangle, \ldots \langle X^L, Y^L \rangle \}$

• Data likelihood $= \prod_l P(X^l, Y^l | W)$

• Data conditional likelihood $= \prod_l P(Y^l | X^l, W)$

$$W \leftarrow \arg \max_W \ln \prod_l P(Y^l | X^l, W)$$
Expressing Conditional Log Likelihood

\[ l(W) \equiv \ln \prod_l P(Y^l | X^l, W) = \sum_l \ln P(Y^l | X^l, W) \]

\[ P(Y = 0 | X, W) = \frac{1}{1 + \exp(w_0 + \sum_i w_i X_i)} \]

\[ P(Y = 1 | X, W) = \frac{\exp(w_0 + \sum_i w_i X_i)}{1 + \exp(w_0 + \sum_i w_i X_i)} \]

\[ l(W) = \sum_l Y^l \ln P(Y^l = 1 | X^l, W) + (1 - Y^l) \ln P(Y^l = 0 | X^l, W) \]

\[ = \sum_l Y^l \ln \frac{P(Y^l = 1 | X^l, W)}{P(Y^l = 0 | X^l, W)} + \ln P(Y^l = 0 | X^l, W) \]

\[ = \sum_l Y^l (w_0 + \sum_i w_i X_i^l) - \ln (1 + \exp(w_0 + \sum_i w_i X_i^l)) \]
Maximizing Conditional Log Likelihood

\[
P(Y = 0|X, W) = \frac{1}{1 + \exp(w_0 + \sum_i w_i X_i)}
\]

\[
P(Y = 1|X, W) = \frac{\exp(w_0 + \sum_i w_i X_i)}{1 + \exp(w_0 + \sum_i w_i X_i)}
\]

\[
l(W) \equiv \ln \prod_l P(Y^l|X^l, W)
\]

\[
= \sum_l Y^l (w_0 + \sum_i^n w_i X_i^l) - \ln(1 + \exp(w_0 + \sum_i^n w_i X_i^l))
\]

Good news: \( l(W) \) is concave function of \( W \)

Bad news: no closed-form solution to maximize \( l(W) \)
Gradient Descent

\[
\nabla E[\bar{w}] \equiv \left[ \frac{\partial E}{\partial w_0}, \frac{\partial E}{\partial w_1}, \ldots, \frac{\partial E}{\partial w_n} \right]
\]

Training rule:

\[
\Delta \bar{w} = -\eta \nabla E[\bar{w}]
\]

i.e.,

\[
\Delta w_i = -\eta \frac{\partial E}{\partial w_i}
\]
Maximize Conditional Log Likelihood: Gradient Ascent

\[ l(W) \equiv \ln \prod_{l} P(Y^l|X^l, W) \]

\[ = \sum_{l} Y^l(w_0 + \sum_{i} w_{i}X_{i}^l) - \ln(1 + e^{\exp(w_0 + \sum_{i} w_{i}X_{i}^l)}) \]

\[ \frac{\partial l(W)}{\partial w_{i}} = \sum_{l} X_{i}^l(Y^l - \hat{P}(Y^l = 1|X^l, W)) \]

Gradient ascent algorithm: iterate until change < \( \varepsilon \)

For all \( i \),  
\[ w_{i} \leftarrow w_{i} + \eta \sum_{l} X_{i}^l(Y^l - \hat{P}(Y^l = 1|X^l, W)) \]

repeat
That’s all M(C)LE. How about MAP?

- One common approach is to define priors on \( W \)
  - Normal distribution, zero mean, identity covariance
- Helps avoid very large weights and overfitting
- MAP estimate

\[
W \leftarrow \arg \max_W \ln P(W|\{\langle Y^l, X^l \rangle \})
\]

\[
W \leftarrow \arg \max_W P(W) \ln \prod_l P(Y^l|X^l, W)
\]
MLE vs MAP

• Maximum conditional likelihood estimate

\[ W \leftarrow \arg \max_W \ln \prod_l P(Y^l|X^l, W) \]

\[ w_i \leftarrow w_i + \eta \sum_l X_i^l (Y^l - \hat{P}(Y^l = 1|X^l, W)) \]

• Maximum a posteriori estimate

\[ W \leftarrow \arg \max_W P(W) \ln \prod_l P(Y^l|X^l, W) \]

\[ w_i \leftarrow w_i - \eta \lambda w_i + \eta \sum_l X_i^l (Y^l - \hat{P}(Y^l = 1|X^l, W)) \]
Naïve Bayes vs. Logistic Regression

• Generative and Discriminative classifiers
• Asymptotic comparison (# training examples → infinity)
  • when model correct
  • when model incorrect
• Non-asymptotic analysis
  • convergence rate of parameter estimates
  • convergence rate of expected error
• Experimental results

[Ng & Jordan, 2002]
Naïve Bayes vs Logistic Regression

Consider $Y$ and $X_i$ boolean, $X=\langle X_1 \ldots X_n \rangle$

Number of parameters:
- NB: $2n +1$
- LR: $n+1$

Estimation method:
- NB parameter estimates are uncoupled
- LR parameter estimates are coupled
What is the difference asymptotically?

Notation: let $\epsilon(h_{A,m})$ denote error of hypothesis learned via algorithm $A$, from $m$ examples

- If assumed naïve Bayes model correct, then

$$\epsilon(h_{Dis,\infty}) = \epsilon(h_{Gen,\infty})$$

- If assumed model incorrect

$$\epsilon(h_{Dis,\infty}) \leq \epsilon(h_{Gen,\infty})$$

Note assumed discriminative model can be correct even when generative model incorrect, but not vice versa.
Rate of convergence: logistic regression

Let $h_{\text{Dis},m}$ be logistic regression trained on $m$ examples in $n$ dimensions. Then with high probability:

$$\epsilon(h_{\text{Dis},m}) \leq \epsilon(h_{\text{Dis},\infty}) + O(\sqrt{\frac{n}{m}} \log \frac{m}{n})$$

Implication: if we want $\epsilon(h_{\text{Dis},m}) \leq \epsilon(h_{\text{Dis},\infty}) + \epsilon_0$ for some constant $\epsilon_0$, it suffices to pick $m = \Omega(n)$

$\Rightarrow$ Convergences to its classifier, in order of $n$ examples

(result follows from Vapnik’s structural risk bound, plus fact that VCDim of $n$ dimensional linear separators is $n$)
Rate of convergence: naïve Bayes

Consider first how quickly parameter estimates converge toward their asymptotic values.

Then we’ll ask how this influences rate of convergence toward asymptotic classification error.
Rate of convergence: naïve Bayes parameters

Let any $\epsilon_1, \delta > 0$ and any $l \geq 0$ be fixed. Assume that for some fixed $\rho_0 > 0$, we have that $\rho_0 \leq p(y = T) \leq 1 - \rho_0$. Let $m = O((1/\epsilon_1^2) \log(n/\delta))$. Then with probability at least $1 - \delta$, after $m$ examples:

1. For discrete inputs, $|\hat{p}(x_i|y = b) - p(x_i|y = b)| \leq \epsilon_1$, and $|\hat{p}(y = b) - p(y = b)| \leq \epsilon_1$, for all $i$, $b$.

2. For continuous inputs, $|\hat{\mu}_i|_{y=b} - \mu_i|_{y=b}| \leq \epsilon_1$, and $|\hat{\sigma}_i^2 - \sigma_i^2| \leq \epsilon_1$, for all $i$, $b$. 
Some experiments from UCI data sets

Figure 1: Results of 15 experiments on datasets from the UCI Machine Learning repository. Plots are of generalization error vs. $m$ (averaged over 1000 random train/test splits). Dashed line is logistic regression; solid line is naive Bayes.
What you should know:

- Logistic regression
  - Functional form follows from Naïve Bayes assumptions
  - But training procedure picks parameters without the conditional independence assumption
  - MLE training: pick $W$ to maximize $P(Y | X, W)$
  - MAP training: pick $W$ to maximize $P(W | X,Y)$
    - ‘regularization’

- Gradient ascent/descent
  - General approach when closed-form solutions unavailable

- Generative vs. Discriminative classifiers
  - Bias vs. variance tradeoff