10-601 Assignment Homework 2

Tanay Gavankar
tgavanka@andrew.cmu.edu
Due: 9/28/12

1.1: Naive Bayes: Basic Concepts

(a) Yes, this is the assumption that Naive Bayes makes.

(b) No. Since all we know that X is independent of $Y|Z$, we do not know if X is independent of Y in general.

(c) $J * (X^n)$. (Note that since this covers the entire space, we'd only have to calculate $J * (X^n) - 1$ of them, as the remaining one would be $1 - \sum P(X_i|Y)$.)

(d) There are n distinct μ_{ij}, σ_{ij} because there is one for each x_i.

(e) When estimating Y, the denominator does not depend on what we are trying to calculator the probability for, so the denominator is effectively constant.

(f) Yes, we can calculate P(X) from the parameters estimated by Naive Bayes by using relative frequencies of the training sett.

1.2: Naive Bayes: Parameter elimination

(a) $\hat{\theta_{1k}} = \frac{\sum_1^M x_{1j}}{M}$

(b)

$$
\begin{aligned}
\mu^{mle} &= argmax_\mu P(X_1, X_2, ...X_n|Y) \\
&= argmax_\mu \prod_{i=1}^{n} P(X_i|Y) \\
&= argmax_\mu \sum_{i=1}^{n} log(P(X_i|Y)) \\
&= argmax_\mu \sum_{i=1}^{n} log(\frac{1}{\sigma\sqrt{2\pi}} exp(\frac{-(X_i - \mu_i)^2}{2\sigma^2})) \\
&= argmax_\mu \frac{1}{\sigma\sqrt{2\pi}} \sum_{i=1}^{n} \frac{-(X_i - \mu_i)^2}{2\sigma^2} \\
&= argmin_\mu \sum_{i=1}^{n} (X_i - \mu_i)^2
\end{aligned}
$$

We want to minimize this, so take the derivative and set to zero, and solve.

$$\begin{aligned} 0 &= \frac{\partial}{\partial \mu} \sum_{i=1}^{n} (X_i - \mu_i)^2 \\ &= -\sum_{i=1}^{n} 2(X_i - \mu) \\ \Rightarrow \mu^{mle} &= \frac{1}{n} \sum_{i=1}^{n} X_i \end{aligned}$$

2: Regularized Multi-Class Logistic Regression

(a) Let l be the lth training data entry.

$$\begin{aligned} W &\leftarrow argmax_W \prod_{l} P(Y^l = k|X^l, W) \\ l(W) &= \sum_{l \in D} lnP(Y^l|X^l, W) \\ l(W) &= \sum_{l \in D} ln(\frac{exp(w_k^T x)}{1 + \sum_{t=1}^{K-1} exp(w_t^T x)}) \\ l(W) &= \sum_{l \in D} ln(exp(w_k^T x)) - ln(1 + \sum_{t=1}^{K-1} exp(w_t^T x))) \\ l(W) &= \sum_{l \in D} w_k^T x - ln(1 + \sum_{t=1}^{K-1} exp(w_t^T x))) \end{aligned}$$

(b)

(c)

(d) Yes, it will converge to a global maximum because that is where the MLE appears.

3: Generative-Discriminative Classifiers

(a) Since X_i are boolean variables, we can use a single parameter to define $P(X_i|Y = y_k)$. Let $\theta_{i1} = P(X_i = 1|Y = 1)$. This means that $P(X_i = 0|Y = 1) = (1 - \theta_{i1})$ and $P(X_i = 1|Y = 0) = \theta_{i0}$. This all means that $P(X_i|Y = 1) = \theta_{i1}^{X_i}(1 - \theta_{i1})^{(1-X_i)}$. Also, let $P(Y = 1) = \pi$.

$$\begin{aligned}
P(Y=1|X) &= \frac{P(Y=1)P(X|Y=1)}{P(Y=1)P(X|Y=1)+P(Y=0)P(X|Y=0)} \\
&= \frac{1}{1+\frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}} \\
&= \frac{1}{1+exp(ln(\frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}))} \\
&= \frac{1}{1+exp(ln\frac{P(Y=0)}{P(Y=1)}+\sum_i ln\frac{P(X_i|Y=0)}{P(X_i|Y=1)})} \\
&= \frac{1}{1+exp(ln\frac{1-\pi}{\pi}+\sum_i ln\frac{\theta_{i0}^{X_i}(1-\theta_{i0})^{(1-X_i)}}{\theta_{i1}^{X_i}(1-\theta_{i1})^{(1-X_i)}})}
\end{aligned}$$

Let's look at only the sum in the denominator.

$$\begin{aligned}
\sum_i ln\frac{P(X_i|Y=0)}{P(X_i|Y=1)} &= \sum_i ln\frac{\theta_{i0}^{X_i}(1-\theta_{i0})^{(1-X_i)}}{\theta_{i1}^{X_i}(1-\theta_{i1})^{(1-X_i)}} \\
&= \sum_i ln(\theta_{i0}^{X_i}(1-\theta_{i0})^{(1-X_i)}) - ln(\theta_{i1}^{X_i}(1-\theta_{i1})^{(1-X_i)}) \\
&= \sum_i ln(\theta_{i0}^{X_i}) + ln((1-\theta_{i0})^{(1-X_i)}) - ln(\theta_{i1}^{X_i}) - ln((1-\theta_{i1})^{(1-X_i)}) \\
&= \sum_i X_i ln(\theta_{i0}) + (1-X_i)ln(1-\theta_{i0}) - X_i ln(\theta_{i1}) - (1-X_i)ln(1-\theta_{i1}) \\
&= \sum_i X_i ln(\theta_{i0}) + ln(1-\theta_{i0}) - X_i ln(1-\theta_{i0}) - X_i ln(\theta_{i1}) - ln(1-\theta_{i1}) + X_i ln(1-\theta_{i1})
\end{aligned}$$

(b) Assuming all the NB assumptions are satisfied, both NB and LR will have identical results because NB can be mapped to LR.

(c) Assuming the conditional independence assumption of NB is not satisfied, then the NB bias will cause it to perform less accurately than LR in the limit.

(d) It is not possible for the LR estimated parameters to calculated P(X) because LR calculates P(Y—X), and so since X has to be given, a probability cannot be calculated for it.