Homework 2
10-601 Machine Learning
Name: Shashank Singh
Email: sss1@andrew.cmu.edu
Due: Friday, September 28, 2012

1 Naive Bayes

Problem 1. Basic Concepts.

a. Yes;

$$\mathsf{P}(X,Y|Z) = \mathsf{P}(X|Y,Z) \cdot \mathsf{P}(Y|Z) = \mathsf{P}(X|Z) \cdot \mathsf{P}(Y|Z).$$

b. Suppose $X = Y = Z$, where $Z \sim \text{Bernoulli}\left(\frac{1}{2}\right)$. Then, $\mathsf{P}(X|Y,Z) = \mathsf{P}(X|Z)$, but

$$\mathsf{P}(X=1,Y=1) = \tfrac{1}{2} \neq \tfrac{1}{4} = \mathsf{P}(X=1) \cdot \mathsf{P}(Y=1).$$

c. Since no θ_{ij} parameter can be determined from any subset of the other θ_{ij} parameters, there are $\boxed{nJ}$ independent θ_{ij} parameters.

d. Since no μ_{ij} or σ_{ij} parameter can be determined from any subset of the other μ_{ij} or σ_{ij} parameters, there are $\boxed{2nJ}$ independent μ_{ij} or σ_{ij} parameters.

e. Since the term $\sum_j \mathsf{P}(Y=y_j) \prod_i \mathsf{P}(X_i|Y=y_j)$ does not depend on y_k,

$$y^* = \text{argmax}_{y_k} \frac{\mathsf{P}(y=y_k) \prod_i \mathsf{P}(x_i|y=y_k)}{\sum_j \mathsf{P}(Y=y_j) \prod_i \mathsf{P}(X_i|Y=y_j)} = \text{argmax}_{y_k} \mathsf{P}(y=y_k) \prod_i \mathsf{P}(x_i|y=y_k).$$

f. Yes; since Naive Bayes is a generative classifier, an estimates of $\mathsf{P}(X)$ can be computed from the parameters estimated by Naive Bayes using Bayes Rule.

Problem 2. Parameter estimation for Naive Bayes

a.

$$\text{MLE}(\hat{\theta}_{1k}) = \boxed{\frac{\sum_{j=1}^{M} x_{1j}}{M}.}$$

b. Since the training instances are independent,

$$P(X|Y) = \prod_{j=1}^{M} P(X_j|Y_j).$$

Thus,

$$\begin{aligned}
\mathrm{MLE}(\mu_{ik}) &= \mathrm{argmax}_{\mu\in\mathbb{R}} \ln\left(\prod_{j=1}^{M} \frac{1}{\sigma_{ik}\sqrt{2\pi}} \exp\left(\frac{-(x_{ij}-\mu_{ik})^2}{2\sigma_{ik}^2}\right)\right) \\
&= \mathrm{argmax}_{\mu\in\mathbb{R}} \sum_{j=1}^{M} \ln\left(\frac{1}{\sigma_{ik}\sqrt{2\pi}} \exp\left(\frac{-(x_{ij}-\mu_{ik})^2}{2\sigma_{ik}^2}\right)\right) \\
&= \mathrm{argmax}_{\mu\in\mathbb{R}} \sum_{j=1}^{M} \ln\left(\exp\left(-(x_{ij}-\mu_{ik})^2\right)\right) \\
&= \mathrm{argmax}_{\mu\in\mathbb{R}} \sum_{j=1}^{M} -(x_{ij}-\mu_{ik})^2.
\end{aligned}$$

Therefore, μ_{ik} is the value which minimizes the sum of squared errors of the x_{ij} values, so that

$$\mu_{ik} = \boxed{\frac{\sum_{j=1}^{M} x_{ij}}{M},}$$

the mean of the x_{ij} values.

2 Regularized Multi-Class Logistic Regression

a. If we let $\mathbf{w}_K = 0$, then, since $\exp(0) = 1$,

$$\begin{aligned}
L(\mathbf{w}_1,\ldots,\mathbf{w}_K) &= \ln\left(\prod_{l=1}^{D} \mathsf{P}\left(Y^l|X^l,W\right)\right) - \sum_{k=1}^{K}\frac{\lambda}{2}\|\mathbf{w}_k\|^2 \\
&= \sum_{l=1}^{D} \ln(\mathsf{P}\left(Y^l|X^l,W\right)) - \sum_{k=1}^{K}\frac{\lambda}{2}\|\mathbf{w}_k\|^2 \\
&= \ln\left(\frac{1}{1+\sum_{t=1}^{K-1}\exp(\mathbf{w}_t\cdot x_t)}\right) \\
&+ \sum_{l=1}^{D} \ln\left(\frac{\exp(\mathbf{w}_k\cdot x_k)}{1+\sum_{t=1}^{K-1}\exp(\mathbf{w}_t\cdot x_t)}\right) - \sum_{k=1}^{K}\frac{\lambda}{2}\|\mathbf{w}_k\|^2 \\
&= \sum_{l=1}^{D} \ln\left(\frac{\exp(\mathbf{w}_k\cdot x_k)}{\sum_{t=1}^{K-1}\exp(\mathbf{w}_t\cdot x_t)}\right) - \sum_{k=1}^{K}\frac{\lambda}{2}\|\mathbf{w}_k\|^2 \\
&= \sum_{l=1}^{D} \mathbf{w}_{Y^l}\cdot x_k - \ln\left[\sum_{t=1}^{K}\exp(\mathbf{w}_t\cdot x_t)\right] - \sum_{k=1}^{K}\frac{\lambda}{2}\|\mathbf{w}_k\|^2.
\end{aligned}$$

b. Differentiating the above expression with respect to w_{ij} gives

$$\begin{aligned}\frac{\partial L(\mathbf{w}_1,\ldots,\mathbf{w}_K)}{\partial w_{ij}} &= \frac{\partial}{\partial w_{ij}}\sum_{l=1}^{D}\mathbf{w}_{Y^l}\cdot x_k - \ln\left[\sum_{t=1}^{K}\exp(\mathbf{w}_t\cdot x_t)\right] - \sum_{k=1}^{K}\frac{\lambda}{2}\|\mathbf{w}_k\|^2.\\ &= \sum_{l=1}^{D}\delta(Y^l=k)x_k^l) - \left[\frac{\partial}{\partial w_{ij}}\ln\left[\sum_{t=1}^{K}\exp(\mathbf{w}_t\cdot x_t)\right]\right] - \lambda w_{ij}\\ &= \left(\sum_{l-1}^{D}X_i^l\left(\delta(Y^l=k) - \mathsf{P}\left(Y^l=k|X^l,\mathbf{w}_1,\ldots,\mathbf{w}_k\right)\right)\right) - \lambda w_{ki}.\end{aligned}$$

where $\delta(Y^l = k)$ is 1 if $Y^l = k$ and 0 otherwise.

Thus, the desired gradient is

$$\frac{\partial L(\mathbf{w}_1,\ldots,\mathbf{w}_K)}{\partial \mathbf{w}_i} = \begin{bmatrix}\frac{\partial L(\mathbf{w}_1,\ldots,\mathbf{w}_K)}{\partial w_{i1}}\\ \frac{\partial L(\mathbf{w}_1,\ldots,\mathbf{w}_K)}{\partial w_{i2}}\\ \vdots\\ \frac{\partial L(\mathbf{w}_1,\ldots,\mathbf{w}_K)}{\partial w_{iK}}\end{bmatrix}$$

c. The update rule is

$$w_{ki} \leftarrow w_{ki} + \nu\left(\sum_{l-1}^{D}X_i^l\left(\delta(Y^l=k) - \mathsf{P}\left(Y^l=k|X^l,\mathbf{w}_1,\ldots,\mathbf{w}_k\right)\right)\right) - \nu\lambda w_{ki}.$$

d. Since the log likelihood function is concave in each $\mathbf{w}_i$, it is concave in $(\mathbf{w}_1,\ldots,\mathbf{w}_k)$, so that the gradient ascent will converge on a global maximum.

3 Generative-Discriminative Classifiers

a. By Bayes Rule and the Law of Total Probability,

$$\begin{aligned}\mathsf{P}(Y=1|X) &= \frac{\mathsf{P}(Y=1)\cdot\mathsf{P}(X|Y=1)}{\mathsf{P}(Y=1)\cdot\mathsf{P}(X|Y=1)+\mathsf{P}(Y=0)\cdot\mathsf{P}(X|Y=0)}\\ &= \frac{1}{1+\frac{\mathsf{P}(Y=0)\cdot\mathsf{P}(X|Y=0)}{\mathsf{P}(Y=1)\cdot\mathsf{P}(X|Y=1)}} = \frac{1}{1+\exp\ln\frac{\mathsf{P}(Y=0)\cdot\mathsf{P}(X|Y=0)}{\mathsf{P}(Y=1)\cdot\mathsf{P}(X|Y=1)}}\\ &= \frac{1}{1+\exp\left(\ln\frac{\mathsf{P}(Y=0)}{\mathsf{P}(Y=1)}+\sum_{i=1}^{n}\ln\frac{\mathsf{P}(X_i|Y=0)}{\mathsf{P}(X_1|Y=1)}\right)}\\ &= \frac{1}{1+\exp\left(\ln\frac{1-\pi}{\pi}+\sum_{i=1}^{n}\ln\frac{\mathsf{P}(X_i|Y=0)}{\mathsf{P}(X_i|Y=1)}\right)}\end{aligned}$$

where $\pi := \mathsf{P}(Y=1)$. Letting $\theta_{ik} := \mathsf{P}(X_i = 1|Y=k)$ (for $k \in \{0,1\}$),

$$\sum_{i=1}^{n} \ln \frac{\mathsf{P}(X_i|Y=0)}{\mathsf{P}(X_i|Y=1)} = \sum_{i=1}^{n} \ln \frac{\theta_{i0}^{X_i}(1-\theta_{i0})^{1-X_i}}{\theta_{i1}^{X_i}(1-\theta_{i1})^{1-X_i}}$$
$$= \sum_{i=1}^{n} \left(\ln \frac{\theta_{i0}(1-\theta_{i1})}{\theta_{i1}(1-\theta_{i0})} \right) X_i + \ln \frac{(1-\theta_{i0})}{(1-\theta_{i1})}.$$

Thus, for

$$w_0 = \ln \frac{1-\pi}{\pi} + \sum_{i=1}^{n} \ln \left(\frac{1-\theta_{i0}}{1-\theta_{i1}} \right) \quad \text{and} \quad w_i = \left(\ln \frac{\theta_{i0}(1-\theta_{i1})}{\theta_{i1}(1-\theta_{i0})} \right) \text{ (for } i \in \{1,2,\ldots,n\}\text{)},$$

$$\mathsf{P}(Y=1|X) = \frac{1}{1+\exp\left(w_0 + \sum_{i=1}^{n} w_i X_i\right)},$$

which is the desired logistic form. ■

b. When the conditional independence assumption holds, Naive Bayes and Logistic Regression are equivalent, so neither produces better results.

c. Since Naive Bayes depends on the assumption that $\mathsf{P}(X_i|Y, X_j) = \mathsf{P}(X_i|X)$ to simplify the parameters it must estimate, it is less accurate than logistic regression when this assumption does not hold.

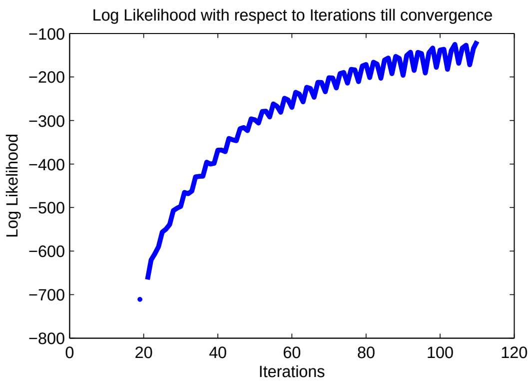
d. No; since Logistic Regression is a discriminative classifier, it cannot estimate $\mathsf{P}(X)$.

4 Programming

4.1 Feature selection with Mutual Information

a.

Classifer	Training Accuracy	Testing Accuracy	Training Time
Logistic Regression	0.9918	0.9878	0.1414 seconds
Naive Bayes	0.9901	0.9834	0.0891 seconds


Log Likelihood with respect to Iterations till convergence
−100
−200
−300
−400
−500
−600
−700
−800
0
20
40
60
80
100
120
Iterations
Log Likelihood