CSE – 426 Homework 3

Griffin Kent

1) Let $\{y^{(i)}\}_{i=1}^{m}$ be an I.I.D. sample from a multinomial distribution with unknown parameters $(\phi_1, \phi_2, \dots, \phi_k)$, where $y^{(i)} \in \{1, \dots, k\}$ and ϕ_j is the probability that a sample is equal to j. Find the MLE of $(\phi_1, \phi_2, \dots, \phi_k)$ from the observations. Your answer should include a log-likelihood function, the partial derivatives of the log-likelihood with respect to ϕ_j, $j = 1, \dots, k$, and then use the constraint that $\sum_j \phi_j = 1$ to find a formula for the MLE of ϕ_j in terms of the observations. [*Hints: one way to solve for ϕ_j, $j = 1, \dots, k$, is to replace ϕ_k with $1 - \sum_{j \neq k} \phi_j$ and maximize the log-likelihood involving $\phi_1, \phi_2, \dots, \phi_{k-1}$ as an unconstrained optimization problem. The other approach is to maximize the log-likelihood with respect to all the k parameters, subject to the constraint that $\sum_j \phi_j = 1$. The Lagrangian method will then be used.*] (Refer to PRML (2.32))

(Proof): To start, we can see that the probability distribution of y is defined as

$$P(y|\boldsymbol{\phi}) = \prod_{j=1}^{k} \phi_j^{1[y=j]}.$$

Now, we can derive the likelihood function

$$L\left(\boldsymbol{\phi}; \{y^{(i)}\}_{i=1}^{m}\right) = \prod_{i=1}^{m} P\left(y^{(i)}|\boldsymbol{\phi}\right) = \prod_{i=1}^{m} \prod_{j=1}^{k} \phi_j^{1[y^{(i)}=j]}$$

$$= \prod_{j=1}^{k} \phi_j^{\sum_{i=1}^{m} 1[y^{(i)}=j]}.$$

Likewise, we can derive the log-likelihood function to be

$$\ell(\boldsymbol{\phi}) = \log L(\boldsymbol{\phi}) = \sum_{j=1}^{k} \log \phi_j \sum_{i=1}^{m} 1[y^{(i)} = j].$$

Setting up a Lagrangian, we have

$$\mathcal{L}(\boldsymbol{\phi}, \lambda) = \sum_{j=1}^{k} \log \phi_j \sum_{i=1}^{m} 1[y^{(i)} = j] + \lambda\left(\sum_{j=1}^{k} \phi_j - 1\right).$$

In order to maximize this function, we will take the derivative with respect to ϕ_j and set it to zero:

$$\frac{\partial \mathcal{L}(\boldsymbol{\phi}, \lambda)}{\partial \phi_j} = \frac{1}{\phi_j} \sum_{i=1}^{m} 1[y^{(i)} = j] + \lambda = 0$$

$$\rightarrow \phi_j = -\frac{\sum_{i=1}^{m} 1\left[y^{(i)} = j\right]}{\lambda}.$$

Substituting this into the constraint, we have

$$\sum_{j=1}^{k} \phi_j = 1 \longrightarrow -\frac{1}{\lambda}\sum_{j=1}^{k}\sum_{i=1}^{m} 1\left[y^{(i)} = j\right] = 1$$

$$\longrightarrow \lambda = -\sum_{j=1}^{k}\sum_{i=1}^{m} 1\left[y^{(i)} = j\right] = -m.$$

Finally, substituting this back into the above equation, we have

$$\phi_j = \frac{\sum_{i=1}^{m} 1\left[y^{(i)} = j\right]}{m}.$$

■

2) Let $y \in \{1, \dots, k\}$ be distributed according to a multinomial distribution with parameters $(\phi_1, \phi_2, \dots, \phi_k)$, where the probability of y taking j is ϕ_j. Let m observations be $\{y^{(1)}, \dots, y^{(m)}\}$, where $y^{(i)}$ is sampled from the multinomial. We know that MLE will estimate ϕ_j as $\sum_{i=1}^{m} \mathbb{1}\left[y^{(i)} = j\right]/m$. Prove that the Laplacian smoothing for multinomial distributions does generate a probability distribution over the set $\{1, \dots, k\}$.
[*Hints: add a hallucinated sample to each class and follow the MLE of multinomial. You can use the conclusion about* ϕ_j^{MLE} *without proving it.*]

(Proof): We are given the MLE of ϕ_j as

$$\phi_j^{MLE} = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\left[y^{(i)} = j\right].$$

Laplacian smoothing will add 1 to the quantity $\sum_{i=1}^{m} \mathbb{1}\left[y^{(i)} = j\right]$, increasing m to $m + k$, where k is the number of classes. Thus, the Laplacian smoothing estimate of the MLE of ϕ_j is defined as

$$\phi_j^* = \frac{1}{m+k}\sum_{i=1}^{m} \mathbb{1}\left[y^{(i)} = j\right] + 1.$$

We can see that

$$\sum_{j=1}^{k} \phi_j^* = \frac{1}{m+k}\sum_{j=1}^{k}\sum_{i=1}^{m} \mathbb{1}\left[y^{(i)} = j\right] + 1 = \frac{m+k}{m+k} = 1.$$

And since we know that $\phi_j^* > 0, \forall j \in \{1, \dots, k\}$, we know that the Laplacian smoothing for the multinomial distribution does indeed generate a probability distribution over the set $\{1, \dots, k\}$.■

3) Let the training data be $\boldsymbol{x}^{(1)} = [1,0,1]^T$, $y^{(1)} = 1$ and $\boldsymbol{x}^{(2)} = [0,1,1]^T$, $y^{(2)} = 0$. Assume the features and labels are all binary. First identify the parameters that need to be estimated for the Naïve Bayes classifier, then write down the log-likelihood of the training data. Lastly give an estimation of the parameters.

(Solution): Given a binary input vector $\boldsymbol{x}$, the Naïve Bayes model has the joint probability density

$$\begin{aligned} P(\boldsymbol{x}, y) &= P(y)P(\boldsymbol{x}_1|y)P(\boldsymbol{x}_2|y)\ldots P(\boldsymbol{x}_n|y) \\ &= P(y)\prod_{j=1}^{n} P(\boldsymbol{x}_j|y). \end{aligned}$$

Thus, each $P(\boldsymbol{x}_j|y)$ is a Bernoulli random variable parameterized by ϕ_{jy}. Then, given a sample $\{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{m}$, the likelihood function is defined as

$$L\left(\boldsymbol{\phi}; \{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{m}\right) = \prod_{i=1}^{m}\prod_{j=1}^{n} P\big(y^{(i)}\big)P(\boldsymbol{x}_j^{(i)}|y^{(i)})$$

Which yields the corresponding log-likelihood function

$$\ell(\boldsymbol{\phi}) = \sum_{i=1}^{m}\sum_{j=1}^{n}\left\{\log P\big(y^{(i)}\big) + \log P\left(\boldsymbol{x}_j^{(i)}\middle|y^{(i)}\right)\right\}.$$

For this problem, the likelihood function is

$$L(\boldsymbol{\phi}) = \prod_{i=1}^{2} P\big(y^{(i)}\big)P\left(\boldsymbol{x}_1^{(i)}\middle|y^{(i)}\right)P\left(\boldsymbol{x}_2^{(i)}\middle|y^{(i)}\right)P\left(\boldsymbol{x}_3^{(i)}\middle|y^{(i)}\right)$$

Which yields the corresponding log-likelihood function

$$\begin{aligned} \ell(\boldsymbol{\phi}) &= \log\left(\prod_{i=1}^{2} P\big(y^{(i)}\big)P\left(\boldsymbol{x}_1^{(i)}\middle|y^{(i)}\right)P\left(\boldsymbol{x}_2^{(i)}\middle|y^{(i)}\right)P\left(\boldsymbol{x}_3^{(i)}\middle|y^{(i)}\right)\right) \\ &= \sum_{i=1}^{2}\log\left(P\big(y^{(i)}\big)P\left(\boldsymbol{x}_1^{(i)}\middle|y^{(i)}\right)P\left(\boldsymbol{x}_2^{(i)}\middle|y^{(i)}\right)P\left(\boldsymbol{x}_3^{(i)}\middle|y^{(i)}\right)\right) \\ &= \sum_{i=1}^{2}\log P\big(y^{(i)}\big) + \log P\left(\boldsymbol{x}_1^{(i)}\middle|y^{(i)}\right) + \log P\left(\boldsymbol{x}_2^{(i)}\middle|y^{(i)}\right) + \log P\left(\boldsymbol{x}_3^{(i)}\middle|y^{(i)}\right). \end{aligned}$$

Since each of these probabilities are defined over a Bernoulli distribution such that $P(y) = \phi_1^{y}(1-\phi_1)^{1-y}$ and $P(x_j|y) = \phi_{j,y}^{y}\big(1-\phi_{j,y}\big)^{1-y}$, where $\phi_1 = P(Y=1)$ and $\phi_{j,y} = P(x_j = 1|Y = y)$. Then we can see that the log likelihood becomes:

$$\begin{aligned} \ell(\boldsymbol{\phi}) = \sum_{i=1}^{2}\Bigg\{&\log\left[\phi_1^{y^{(i)}}(1-\phi_1)^{1-y^{(i)}}\right] + \log\left[\phi_{1,y^{(i)}}^{y^{(i)}}\left(1-\phi_{1,y^{(i)}}\right)^{1-y^{(i)}}\right] \\ &+ \log\left[\phi_{2,y^{(i)}}^{y^{(i)}}\left(1-\phi_{2,y^{(i)}}\right)^{1-y^{(i)}}\right] + \log\left[\phi_{3,y^{(i)}}^{y^{(i)}}\left(1-\phi_{3,y^{(i)}}\right)^{1-y^{(i)}}\right]\Bigg\}. \end{aligned}$$

Now, using the MLE of $\phi_y = P(Y=1)$ is $\frac{\sum_i^m 1[y^{(i)}=y]}{m}$, and using the Laplacian smoothing estimate of MLE for ϕ_{jy}, which is $\frac{\sum_i^m 1\left[x_j^{(i)}=1, y^{(i)}=y\right]+1}{\sum_i^m 1[y^{(i)}=y]+1}$, we can obtain the following parameter estimates for our training data:

$$\phi_1 = \frac{1\big[y^{(1)}=1\big] + 1\big[y^{(2)}=1\big]}{2} = \frac{1+0}{2} = \frac{1}{2}.$$

$$\phi_{10} = \frac{1\left[\boldsymbol{x}_1^{(1)} = 1, y^{(1)} = 0\right] + 1\left[\boldsymbol{x}_1^{(2)} = 1, y^{(2)} = 0\right] + 1}{1[y^{(1)} = 0] + 1[y^{(2)} = 0] + 1} = \frac{0+0+1}{0+1+1} = \frac{1}{2}.$$

$$\phi_{11} = \frac{1\left[\boldsymbol{x}_1^{(1)} = 1, y^{(1)} = 1\right] + 1\left[\boldsymbol{x}_1^{(2)} = 1, y^{(2)} = 1\right] + 1}{1[y^{(1)} = 1] + 1[y^{(2)} = 1] + 1} = \frac{1+0+1}{1+0+1} = 1.$$

$$\phi_{20} = \frac{1\left[\boldsymbol{x}_2^{(1)} = 1, y^{(1)} = 0\right] + 1\left[\boldsymbol{x}_2^{(2)} = 1, y^{(2)} = 0\right] + 1}{1[y^{(1)} = 0] + 1[y^{(2)} = 0] + 1} = \frac{0+1+1}{0+1+1} = 1.$$

$$\phi_{21} = \frac{1\left[\boldsymbol{x}_2^{(1)} = 1, y^{(1)} = 1\right] + 1\left[\boldsymbol{x}_2^{(2)} = 1, y^{(2)} = 1\right] + 1}{1[y^{(1)} = 1] + 1[y^{(2)} = 1] + 1} = \frac{0+0+1}{1+0+1} = \frac{1}{2}.$$

$$\phi_{30} = \frac{1\left[\boldsymbol{x}_3^{(1)} = 1, y^{(1)} = 0\right] + 1\left[\boldsymbol{x}_3^{(2)} = 1, y^{(2)} = 0\right] + 1}{1[y^{(1)} = 0] + 1[y^{(2)} = 0] + 1} = \frac{0+1+1}{0+1+1} = 1.$$

$$\phi_{31} = \frac{1\left[\boldsymbol{x}_3^{(1)} = 1, y^{(1)} = 1\right] + 1\left[\boldsymbol{x}_3^{(2)} = 1, y^{(2)} = 1\right] + 1}{1[y^{(1)} = 1] + 1[y^{(2)} = 1] + 1} = \frac{1+0+1}{1+0+1} = 1.$$

■

4) Given two training examples $\boldsymbol{x}^{(1)} = [2,1]^T$, $y^{(1)} = 1$ and $\boldsymbol{x}^{(2)} = [1,-1]^T$, $y^{(2)} = -1$. Find the functional margin and geometric margin of the two training examples to the hyperplane $h(\boldsymbol{x}; \boldsymbol{w}, b) = \boldsymbol{w}^T\boldsymbol{x} + b$ with $\boldsymbol{w} = [1,-1]^T$ and $b = 2$. Compute the normal direction $\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$ and the distance from the origin to the hyperplane $\frac{b}{\|\boldsymbol{w}\|}$. Then draw the two training examples and the hyperplane with its normal direction in $\mathbb{R}^2$.

[*Hints: use the formula in the SVM lecture notes to answer this question and there is no need to derive the formula. For drawing, hand-drawing is acceptable so long as the positions of the asked elements are approximately correct.*]

(Solution): Given a hyperplane $h(\boldsymbol{x}; \boldsymbol{w}, b) = \boldsymbol{w}^T\boldsymbol{x} + b$, the geometric margin of the point $\boldsymbol{x}$ is defined as

$$\gamma = \frac{y(\boldsymbol{w}^T\boldsymbol{x} + b)}{\|\boldsymbol{w}\|_2}.$$

Thus, the geometric margins of the two training examples are:

$$\gamma^{(1)} = \frac{y^{(1)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(1)} + b\right)}{\|\boldsymbol{w}\|_2} = \frac{1(2 + (-1) + 2)}{\sqrt{1+1}} = \frac{3}{\sqrt{2}}.$$

$$\gamma^{(2)} = \frac{y^{(2)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(2)} + b\right)}{\|\boldsymbol{w}\|_2} = \frac{-1(1 + 1 + 2)}{\sqrt{1+1}} = -\frac{4}{\sqrt{2}}.$$

The functional margin of the point $\boldsymbol{x}$ with label y is defined as

$$\hat{\gamma} = y(\boldsymbol{w}^T\boldsymbol{x} + b)$$

Thus, the functional margins of the two training examples are:

$$\hat{\gamma}^{(1)} = y^{(1)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(1)} + b\right) = 1(2 + (-1) + 2) = 3.$$

$$\hat{\gamma}^{(2)} = y^{(2)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(2)} + b\right) = -1(1 + 1 + 2) = -4.$$

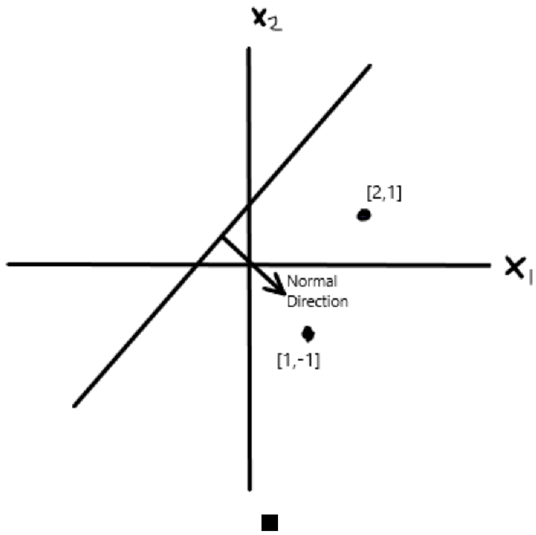
The normal direction is

$$\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}.$$

The distance from the origin to the hyperplane is

$$\frac{b}{\|\boldsymbol{w}\|_2} = \frac{2}{\sqrt{2}} = \sqrt{2}.$$

Here is the visual representation of the hyperplane with its normal direction and the two points:



5) (Graduate Only) Prove that GDA (Gaussian Discriminant Analysis) with two classes and the same Σ for both classes' Gaussians, leads to a logistic regression model. (Essentially, prove Eq. (4.66) and (4.67) in PRML).

(Proof): Recall that a logistic regression model is the parameterized function $h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sigma(\boldsymbol{\theta}^T \boldsymbol{x})$. Likewise, the GDA model uses Bayes rule to calculate the posterior distribution

$$P(y|\boldsymbol{x}) = \frac{P(y)P(\boldsymbol{x}|y)}{P(\boldsymbol{x})}$$

Where $P(\boldsymbol{x}|y)$ is the multivariate gaussian distribution

$$P(\boldsymbol{x}, y = c) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(x-\boldsymbol{\mu}_c)^T\boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu}_c)}.$$

Thus, we can restate our goal: Prove that $P(y|\boldsymbol{x})_{GDA} = \sigma(a)$, where a is some linear function.

To this end,

$$P(y = 1|\boldsymbol{x}) = \frac{P(y = 1)P(\boldsymbol{x}|y = 1)}{P(\boldsymbol{x})}$$

$$= \frac{P(y=1)P(\boldsymbol{x}|y=1)}{P(y=1)P(\boldsymbol{x}|y=1)+P(y=0)P(\boldsymbol{x}|y=0)} = \frac{1}{1+\frac{P(y=0)P(\boldsymbol{x}|y=0)}{P(y=1)P(\boldsymbol{x}|y=1)}}$$

$$= \frac{1}{1+exp\left\{\log\frac{P(y=0)P(\boldsymbol{x}|y=0)}{P(y=1)P(\boldsymbol{x}|y=1)}\right\}}.$$

Expanding out the contents of the exponent,

$$\log\frac{P(y=0)P(\boldsymbol{x}|y=0)}{P(y=1)P(\boldsymbol{x}|y=1)} = \log\frac{P(y=0)}{P(y=1)} + \log\frac{P(\boldsymbol{x}|y=0)}{P(\boldsymbol{x}|y=1)}$$

$$= \log\frac{P(y=0)}{P(y=1)} + \log\frac{\frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_0)}}{\frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)}}$$

$$= \log\frac{P(y=0)}{P(y=1)} + \log\left\{e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_0)+\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)}\right\}$$

$$= \log\frac{P(y=0)}{P(y=1)} + -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_0) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)$$

$$= \log\frac{P(y=0)}{P(y=1)} + \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 - \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \frac{1}{2}\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0$$

$$= \{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)\}^T\boldsymbol{x} + \frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \frac{1}{2}\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 + \log\frac{P(y=0)}{P(y=1)}$$

$$= w^T\boldsymbol{x} + w_0.$$

Here, $w = \{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)\}$ and $w_0 = \frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \frac{1}{2}\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 + \log\frac{P(y=0)}{P(y=1)}$.

Now, inserting this expression back into the following:

$$\frac{1}{1+e^{w^T x+w_0}} = \frac{1}{1+e^{w^T x+w_0}} = \sigma(a).$$

Where, $a = -w^T\boldsymbol{x} - w_0$. Therefore, since a is a linear function, we have proven that GDA with two classes and the same covariance matrix for both classes, yields a logistic regression model.■