EECS 545 Homework 2 Solution (F20)

1. Maximum Likelihood Estimation (5 points)

(a) Answer: $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$
First note the PMF of the Bernoulli distribution with parameter θ:

$$f(x_i, \theta) = \theta^{x_i}(1-\theta)^{(1-x_i)}$$

$$logf(x_i, \theta) = x_i log\theta + (1-x_i)log(1-\theta)$$

Taking the first derivative and setting equal to 0 to find the critical point:

$$\begin{aligned}\frac{\partial}{\partial\theta}\sum_{i=1}^{n} logf(x_i,\theta) &= \sum_{i=1}^{n}\frac{x_i}{\theta} - \frac{1-x_i}{1-\theta}\\ 0 &= \sum_{i=1}^{n}(1-\theta)x_i - \theta(1-x_i)\\ \theta &= \frac{1}{n}\sum_{i=1}^{n} x_i\end{aligned}$$

(b) The parameter θ is a scalar, so the Hessian of the log-likelihood is the second derivative wrt θ:

$$\begin{aligned}\frac{\partial^2}{\partial^2\theta}\sum_{i=1}^{n} logf(x_i,\theta) &= \sum_{i=1}^{n}\frac{x_i}{\theta} - \frac{1-x_i}{1-\theta}\\ &= \frac{-\sum_{i=1}^{n} x_i}{\theta^2} + \frac{-\left(n-\sum_{i=1}^{n} x_i\right)}{1-\theta^2}\end{aligned}$$

Note that $x_i \in \{0,1\}$ and $\theta \in [0,1]$ implies that the numerators in both above terms are negative, and both denominators are positive. Hence, the Hessian is negative for all $\theta \in [0,1]$, and so the critical point found in part (a) is indeed unique, and gives *the* maximum likelihood estimate.

2. Naïve Bayes for Spam Filtering (10 points)

(a) Answer: the additional assumption is that the occurrence of each word in a document is independent. This is stronger than Naive Bayes which only requires *features* to be independent (conditioned on class).

Note that our features x_j are the number of times word j appears in a given document. When we compute $(p_{kj})^(x_j)$ we are treating each occurrence of the word j as an independent event with likelihood p_{kj}. Then $P(X_j = \ell | Y = k) = P(X_j = 1 | Y = k)^\ell = (p_{kj})^\ell$.

(b) Given

$$\hat{y_i} = \arg \max_{k \in \{0,1\}} \log \left(\pi_k \Pi_{j=1}^d {p_{kj}}^{x_{ij}} \right)$$

Distribute the log:

$$\log \left(\pi_k \Pi_{j=1}^d {p_{kj}}^{x_{ij}} \right) = \log \pi_k + \sum_{j=1}^d x_{ij} \log p_{kj}$$

Substituting the definition of p_{kj} gives:

$$\log \left(\pi_k \Pi_{j=1}^d {p_{kj}}^{x_{ij}} \right) = \log \pi_k + \sum_{j=1}^d x_{ij} \left(\log(n_{kj} + \alpha) - \log(n_k + \alpha d) \right)$$

Also correct: the definition of $\pi_k = n_k / n$ may be substituted

$$\log \left(\pi_k \Pi_{j=1}^d {p_{kj}}^{x_{ij}} \right) = \log n_k - \log n + \sum_{j=1}^d x_{ij} \left(\log(n_{kj} + \alpha) - \log(n_k + \alpha d) \right)$$

Optional: the above can be further simplified with vector notation, resulting in a linear classifier of the form

$$\hat{y_i} = \arg \max_{k \in \{0,1\}} b_k + \mathbf{w}_k^T x_i$$

Where $b_k = \log \pi_k$, and $w_{kj} = \log p_{kj} = \log(n_{kj} + \alpha) - \log(n_k + \alpha d)$

(c) $\hat{\pi_0} = 0.4983$, and $\hat{\pi_1} = 0.5017$

(d) The correct test error is 12.5945%, (or an accuracy of 87.41%).

(e) The correct majority-vote predictor always chooses class 1 over class 0, resulting in a test error of 49.8741%, (or an accuracy of 50.13%).

(Note that the answer here is *not* $\hat{\pi_0}$, as $\hat{\pi_0}$ is computed on the training data. The answer 49.8741% is equivalent to the estimate of the class 0 prior on the test data.).

3. **Logistic regression objective function** (5 pts each)

(a) Recall logistic regression is assuming the following likelihood function:

$$\begin{aligned} P(y=1|\tilde{\boldsymbol{x}};\boldsymbol{\theta}) &= \frac{1}{1+e^{-\boldsymbol{\theta}^T\tilde{\boldsymbol{x}}}} \\ P(y=-1|\tilde{\boldsymbol{x}};\boldsymbol{\theta}) &= \frac{e^{-\boldsymbol{\theta}^T\tilde{\boldsymbol{x}}}}{1+e^{-\boldsymbol{\theta}^T\tilde{\boldsymbol{x}}}} \\ &= \frac{1}{1+e^{\boldsymbol{\theta}^T\tilde{\boldsymbol{x}}}} \end{aligned}$$

Alternatively we can write:

$$P(y|\tilde{\boldsymbol{x}};\boldsymbol{\theta}) = \frac{1}{1+e^{-y\boldsymbol{\theta}^T\tilde{\boldsymbol{x}}}}$$

Thus the negative log-likelihood function:

$$-\ell(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \log P(y_i|\tilde{\boldsymbol{x}}_i;\boldsymbol{\theta}) = \sum_{i=1}^{n} \log(1+\exp(-y_i\boldsymbol{\theta}^T\tilde{\boldsymbol{x}}_i)$$

Hence with the new notation of $\phi(t) = \log(1+\exp(-t))$, the logistic regression regularized negative log-likelihood may be written

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{n} \phi(y_i\boldsymbol{\theta}^T\tilde{\boldsymbol{x}}_i) + \lambda\|\boldsymbol{\theta}\|^2.$$

(b) First by chain rule, we have:

$$\nabla_{\boldsymbol{\theta}}\phi(y_i\boldsymbol{\theta}^T\tilde{\boldsymbol{x}}_i) = \phi'(y_i\boldsymbol{\theta}^T\tilde{\boldsymbol{x}}_i)y_i\tilde{\boldsymbol{x}}_i$$

where $\phi'(t) = \frac{-\exp(-t)}{1+\exp(-t)} = -\frac{1}{1+\exp(t)}$
Then by linearity of gradient, we have:

$$\begin{aligned} \nabla J(\boldsymbol{\theta}) &= 2\lambda\boldsymbol{\theta} + \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}}\phi(y_i\boldsymbol{\theta}^T\tilde{\boldsymbol{x}}_i) \\ &= 2\lambda\boldsymbol{\theta} - \sum_{i=1}^{n} y_i\left(\frac{1}{1+\exp(y_i\theta^T\tilde{\boldsymbol{x}}_i)}\right)\tilde{\boldsymbol{x}}_i \end{aligned}$$

Alternatively answer:

$$\nabla J(\boldsymbol{\theta}) = 2\lambda\boldsymbol{\theta} - \sum_{i=1}^{n} y_i\left(\frac{\exp(-y_i\theta^T\tilde{\boldsymbol{x}}_i)}{1+\exp(-y_i\theta^T\tilde{\boldsymbol{x}}_i)}\right)\tilde{\boldsymbol{x}}_i$$

(c) The Hessian

$$\begin{aligned}
\mathbf{H} &= \frac{\partial}{\partial \boldsymbol{\theta}^T}\left(\frac{\partial J(\theta)}{\partial \boldsymbol{\theta}}\right) \\
&= \frac{\partial}{\partial \theta^T}\left\{2\lambda\boldsymbol{\theta} - \sum_{i=1}^{n} y_i\big(\frac{1}{1+\exp(y_i\theta^T\tilde{\boldsymbol{x}}_i)}\big)\tilde{\boldsymbol{x}}_i\right\} \\
&= 2\lambda\mathcal{I} + \sum_{i=1}^{n} \tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i{}^T y_i^2\Big(\frac{\exp(y_i\theta^T\tilde{\boldsymbol{x}}_i)}{\big[1+\exp(y_i\theta^T\tilde{\boldsymbol{x}}_i)\big]^2}\Big) \\
&= 2\lambda\mathcal{I} + \sum_{i=1}^{n} \tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i{}^T \Big(\frac{\exp(y_i\theta^T\tilde{\boldsymbol{x}}_i)}{\big[1+\exp(y_i\theta^T\tilde{\boldsymbol{x}}_i)\big]^2}\Big)
\end{aligned}$$

Note:

$$\begin{aligned}
\frac{\exp(y_i\theta^T\tilde{\boldsymbol{x}}_i)}{\big[1+\exp(y_i\theta^T\tilde{\boldsymbol{x}}_i)\big]^2} &= \frac{1}{\big[1+\exp(y_i\theta^T\tilde{\boldsymbol{x}}_i)\big]\big[1+\exp(-y_i\theta^T\tilde{\boldsymbol{x}}_i)\big]} \\
&= \frac{1}{2+\exp(y_i\theta^T\tilde{\boldsymbol{x}}_i)+\exp(-y_i\theta^T\tilde{\boldsymbol{x}}_i)}
\end{aligned}$$

So any form of above are correct answers.

(d) Letting $a_i = \frac{\exp(y_i\theta^T\tilde{\boldsymbol{x}}_i)}{\big[1+\exp(y_i\theta^T\tilde{\boldsymbol{x}}_i)\big]^2} > 0$ regardless of $\tilde{\boldsymbol{x}}_i$ and y_i, we have for any $\boldsymbol{z} \in \mathbb{R}^d$ such that $\boldsymbol{z} \neq 0$:

$$\begin{aligned}
\boldsymbol{z}^T\mathbf{H}\boldsymbol{z} &= \boldsymbol{z}^T\left(\sum_{i=1}^{n}\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i{}^T a_i + 2\lambda\mathcal{I}\right)\boldsymbol{z} \\
&= \sum_{i=1}^{n} a_i(\boldsymbol{z}^T\tilde{\boldsymbol{x}}_i)(\tilde{\boldsymbol{x}}_i{}^T\boldsymbol{z}) + 2\lambda\boldsymbol{z}^T\boldsymbol{z} \\
&= \textstyle\sum_{i=1}^{n} a_i(\boldsymbol{z}^T\tilde{\boldsymbol{x}}_i)^2 + 2\lambda\|\boldsymbol{z}\|^2
\end{aligned}$$

Observe:
1) when $\lambda \geq 0$, we have $\boldsymbol{z}^T\mathbf{H}\boldsymbol{z} \geq 0, \forall \boldsymbol{z}$ (i.e Hessian is PSD everywhere), hence the problem is convex.
2) when $\lambda > 0$, we have $\boldsymbol{z}^T\mathbf{H}\boldsymbol{z} > 0, \forall \boldsymbol{z} \neq 0$ (i.e Hessian is PD), hence the problem is strictly convex.

4. Logistic Regression for Fashion Classification (15 points)

Test error = 3.2-3.4%
Number of iterations = 8 or 9
Value of objective function after convergence = 456.6390
See 1 for the figure of the misclassified images. We define confidence as the distance to the learned hyperplane. The further a point is away from the hyperplane, the more confident the classifier is.
You can find the solution code for this problem on Canvas.

Figure 1: P4 Figure