CMSC 25025 / STAT 37601
Machine Learning and Large Scale Data Analysis

Assignment 2 Sample Solutions

April 30, 2013

1. 1

$$\mathbb{P}(Y=1) = \mathbb{P}(Y=0) = \frac{1}{2}$$

a The Bayes classifier is of the form

$$h^*(X) = \begin{cases} 1 & \text{if } \mathbb{P}(Y=1|X) \geq \mathbb{P}(Y=0|X) \\ 0 & \text{if } o.w. \end{cases}$$

We have using Bayes theorem that

$$\mathbb{P}(Y|X) = \frac{\mathbb{P}(Y)\mathbb{P}(X|Y)}{\mathbb{P}(X)}$$

Since the other terms are same for both $Y=1$ and $Y=0$ it is easy to see that

$$\frac{\mathbb{P}(Y=1|X)}{\mathbb{P}(Y=0|X)} = \frac{\mathbb{P}(X|Y=1)}{\mathbb{P}(X|Y=0)}$$

$X|Y=0 \sim \mathcal{N}(0,1)$ is a standard Gaussian distribution.
$X|Y=1 \sim \frac{1}{2}\mathcal{N}(-5,1) + \frac{1}{2}\mathcal{N}(5,1)$ is a mixture model i.e.

$$\mathbb{P}(X|Y=1) = \frac{1}{2}\frac{1}{\sqrt{2\pi}}\exp(-\frac{(X+5)^2}{2}) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\exp(-\frac{(X-5)^2}{2})$$

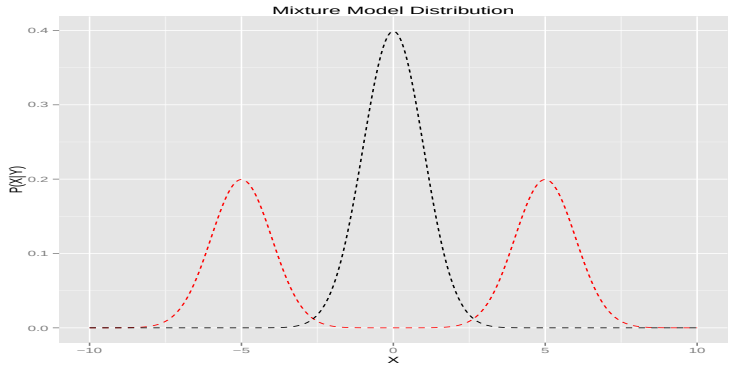
Thus the Bayes classifier chooses label 1 when the second probability is greater than the first. A figure of the distributions is given in figure 1. To figure out the points of intersection, we can take the ratio of the two Gaussian distributions which after some algebra will give the equation

$$\frac{1}{2}\left(\exp(-\frac{1}{2}(10X+25)) + \exp(-\frac{1}{2}(-10X+25))\right) \geq 1$$

We can solve for the points of intersection which gives

$$\frac{1}{\exp(5X)\exp(25/2)} + \frac{\exp(5x)}{\exp(25/2)} = 2$$

Figure 1: Plot of the distributions



Solving for the quadratic on $\exp(5X)$, we get

$$\begin{aligned} \exp(5X) &= \exp(25/2) \pm \sqrt{\exp(25) - 1} \\ X &= \frac{1}{5} \log\left(\exp(25/2) \pm \sqrt{\exp(25) - 1}\right) \\ &= \pm 2.639 \end{aligned}$$

which are the points of intersection. Thus we predict $Y = 1$ for $|X| > 2.639$ and $Y = 0$ otherwise. This is the Bayes classifier.
The Bayes risk is the given by the following expression

$$\begin{aligned} R(h^*) &= \mathbb{P}(Y \neq h^*(X)) \\ &= \int_{-\infty}^{-2.639} \frac{1}{2} \frac{1}{\sqrt{2\pi}} \exp(\frac{-X^2}{2}) dX + \int_{2.639}^{\infty} \frac{1}{2} \frac{1}{\sqrt{2\pi}} \exp(\frac{-X^2}{2}) dX \\ &+ \int_{-2.639}^{2.639} \frac{1}{2} \left(\frac{1}{2} \frac{1}{\sqrt{2\pi}} \exp(-\frac{(X+5)^2}{2}) + \frac{1}{2} \frac{1}{\sqrt{2\pi}} \exp(-\frac{(X-5)^2}{2}) \right) dX \end{aligned}$$

b A linear classifier will be of the form

$$h(X) = \text{sign}(X - \alpha)$$

for some $\alpha \in \mathbb{R}$. You can plug this into the risk expression to get

$$\begin{aligned} R(h) &= P(Y \neq h(X)) \\ &= P(h(X) = 0|Y = 1)P(Y = 1) + P(h(X) = 1|Y = 0)P(Y = 0) \\ &= \int_{-\infty}^{\alpha} \frac{1}{2} \frac{1}{2} \left(\frac{1}{2} \frac{1}{\sqrt{2\pi}} \exp(-\frac{(X+5)^2}{2}) + \frac{1}{2} \frac{1}{\sqrt{2\pi}} \exp(-\frac{(X-5)^2}{2}) \right) dX \\ &+ \int_{\alpha}^{\infty} \frac{1}{2} \frac{1}{\sqrt{2\pi}} \exp(\frac{-X^2}{2}) dX \end{aligned}$$

You can optimize the risk for arbitrary α to get the value of $\alpha = 2.639$ and the corresponding value of Bayes risk.

2 a The Bayes risk is again given by

$$h^*(X) = \begin{cases} 1 & \text{if } \frac{\mathbb{P}(X|Y=1)}{P(X|Y=0)} \geq \frac{1-\pi_1}{\pi_1} \\ -1 & \text{if } o.w. \end{cases}$$

Since $\pi_1 = 1 - \pi_1 = \frac{1}{2}$, the condition is $h^*(\mathbf{x}) = 1$ if $\frac{\mathbb{P}(X|Y=1)}{P(X|Y=0)} > 1$. Since $P(X|Y=1)$ has no support on $X \in [-10, -5)$ (and same for label 0 on $X \in (5, 10]$) it is easy to see what the label will be in those regions.
Since there is uniform probability of $X \in [-5, 5]$ belonging to either of the labels, the risk corresponding to classifying this region as either $+1$ or -1 will be the same. Thus the following is a Bayes classifier

$$h^*(X) = \begin{cases} 1 & \text{if } X > \alpha \forall \alpha \in [-5, 5] \\ -1 & \text{if } o.w. \end{cases}$$

The Bayes risk is given by

$$\begin{aligned} R(h) &= \mathbb{P}(Y \neq h(X)) \\ &= \mathbb{P}(h(X) = 1|Y = -1)\mathbb{P}(Y = -1) + \mathbb{P}(h(X) = -1|Y = 1)\mathbb{P}(Y = 1) \\ &= \frac{5 - \alpha}{5 - (-10)}\frac{1}{2} + \frac{\alpha - (-5)}{10 - (-5)}\frac{1}{2} \\ &= \frac{1}{3} \end{aligned}$$

b A linear classifier is of the form

$$h(x) = sign(x - \alpha)$$

As shown above, for any value of $\alpha \in [-5, 5]$ the risk is $\frac{1}{3}$ which is in fact the Bayes risk. Thus it is also the minimum possible risk for a linear classifier.

c In order to evaluate the risk of the hinge loss, you need to consider various cases.

$$\begin{aligned} \mathbb{E}\left[(1 - Y\beta X)_+\right] &= \mathbb{E}\left[\max(1 - Y\beta X, 0)\right] \\ &= \int \mathbb{P}(Y = 1)\mathbb{P}(\beta X < 1|Y = 1)(1 - \beta X)dP(X) \\ &+ \int \mathbb{P}(Y = -1)\mathbb{P}(\beta X >= 1|Y = -1)(1 + \beta X)dP(X) \\ &= \frac{1}{2}\int_{X<1/\beta|Y=1}(1 - \beta X)dP(X) + \frac{1}{2}\int_{X>-1/\beta|Y=-1}(1 + \beta X)dP(X) \end{aligned}$$

Now we divide β into three possible sets, each of which will give us a different answer. Consider the case $\frac{1}{\beta} \leq 10$ and $\frac{1}{\beta} \geq -5$ which gives $\frac{1}{\beta} \in [-5, 10]$. Then we have the risk

$$R(\text{hinge}) = \frac{1}{2}\int_{-5}^{1/\beta}\frac{1}{15}(1 - \beta X)d(X) + \frac{1}{2}\int_{-\frac{1}{\beta}}^{5}\frac{1}{15}(1 + \beta X)d(X)$$

After some algebra this comes out to

$$\frac{1}{15}\left[\frac{1}{2\beta}+5+12.5\beta\right]$$

If $1/\beta > 10$, the upper bounds of both integrals change and we get

$$\begin{aligned} R(\text{hinge}) &= \frac{1}{2}\int_{-5}^{10}\frac{1}{15}(1-\beta X)d(X) + \frac{1}{2}\int_{-10}^{5}\frac{1}{15}(1+\beta X)d(X) \\ &= 1-\frac{5}{2}\beta \end{aligned}$$

If $1/\beta < -5$ then we have

$$\begin{aligned} R(\text{hinge}) &= \frac{1}{2}\int_{-5}^{1/\beta}\frac{1}{15}(1-\beta X)d(X) + \frac{1}{2}\int_{-1/\beta}^{5}\frac{1}{15}(1+\beta X)d(X) \\ &= 0 \end{aligned}$$

since the upper bound is less than the lower bound. Thus

$$\mathbb{E}\left[(1-Y\beta X)_+\right] = \begin{cases} 0 & \text{if } \frac{1}{\beta} < -5 \\ \frac{1}{15}\left(\frac{1}{2\beta}+5+12.5\beta\right) & \text{if } \frac{1}{\beta} \in [-5,10] \\ 1-\frac{5}{2}\beta & \text{if } \frac{1}{\beta} > 10 \end{cases}$$

2. Logistic Regression

a Please go through the section titled "Fitting Binary Class Logistic Regression" in the class notes for a detailed derivation. The key point is to note that the update

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} - \left(\frac{\delta^2\mathcal{L}(\hat{\beta}^{(k)})}{\delta\beta\delta\beta^\top}\right)^{-1}\frac{\delta\mathcal{L}(\hat{\beta}^{(k)})}{\delta\beta}$$

can be written as

$$\begin{aligned} \hat{\beta}^{(k+1)} &= \hat{\beta}^{(k)} + (X^\top W X)^{-1}X^\top(y-\pi_1^k) \\ &= (X^\top W X)^{-1}X^\top W z^{(k)} \end{aligned}$$

where $z^{(k)} = X\hat{\beta}^{(k)} + W^{-1}(y-\pi_1^{(k)})$ which is the solution obtained by solving the IRLS equations.

b For the labels ± 1 the logistic likelihood is given by

$$\begin{aligned} \mathcal{L}(\hat{\beta}) &= \sum_{i=1}^{n}\left(y_i\mathbf{x}_i^\top\hat{\beta} - \log(1+\exp(\mathbf{x}_i^\top\hat{\beta}))\right) \\ &= \sum_{i=1}^{n}\log\frac{\exp(y_i\mathbf{x}_i^\top\hat{\beta})}{(1+\exp(y_i\mathbf{x}_i^\top\hat{\beta}))} \end{aligned}$$

If the data is separable, there exists a β^* such that

$$\mathbf{x}^\top \beta \begin{cases} \geq 0 & \text{for all } \mathbf{x}_i | y_i = 1 \\ 0 & \text{for all } \mathbf{x}_i | y_i = 0 \end{cases}$$

Thus $y_i \mathbf{x}^\top \beta \geq 0$. Thus when $y_i = 1$ the loglikelihood term is greater than 0 for β^*. Similarly when $y_i = 0$(or -1) the loglikelihood term is greater than 0 as well. As a result, for any $c > 1$, $c\beta^*$ will increase the value of loglikelihood even more and thus the maximum value is reached at infinity. Hence the maximum likelihood estimator of the logistic regression model will not exist and the IRLS algorithm will not converge in this case.

c The Newton algorithm in this case looks like

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} - \left(\frac{\delta^2 \mathcal{L}(\hat{\beta}^{(k)})}{\delta \beta \delta \beta^\top} \right)^{-1} \frac{\delta \mathcal{L}(\hat{\beta}^{(k)})}{\delta \beta}$$

The $\mathcal{L}$ is the regularized loglikehood in this case so that the gradient looks like

$$\begin{aligned} \frac{\delta \mathcal{L}(\hat{\beta}^{(k)})}{\delta \beta} &= -\sum_{i=1}^{n} \left(y_i - \pi_1(\mathbf{x}, \hat{\beta}^{(k)}) \right) \mathbf{x}_i + 2\lambda \hat{\beta}^{(k)} \\ &= -X^\top (\mathbf{y} - \pi_1^{(k)}) + 2\lambda \hat{\beta}^{(k)} \end{aligned}$$

Here $X \in \mathbb{R}^{n \times d}, \mathbf{y}, \pi_1^{(k)} \in \mathbb{R}^n, \hat{\beta}^{(k)} \in \mathbb{R}^d$ so that the gradient $\in \mathbb{R}^d$. The Hessian in this case will look like

$$H = -X^\top W X + 2\lambda I$$

where I is the $d \times d$ identity matrix. If you have incorporated the intercept β_0 in β so that your $\beta \in \mathbb{R}^{d+1}$ you will have a $(d+1) \times (d+1)$ diagonal matrix as the Hessian with a 1 corresponding to all the other variables and a 0 corresponding to the intercept. The rest of the algorithm remains the same.

This algorithm penalizes the squared norm of the model and as a result penalizes the size of the model in some sense. Unlike simple IRLS solution, the solution in this case will be forced to move toward the origin and cannot be scaled arbitrarily as the regularizer $\|\beta\|^2$ will become very large in that case. As a result, the maximum likelihood for this case can always be minimized.