Homework 2

Maxwell Aladago

February 26, 2018

1 q1

1.1 q1a

Given $E^{LWLR}(\theta) = \frac{1}{2}\sum_{i=1}^{m} w^{(i)}(x)[y^{(i)} - \theta^T b^l(x^{(i)})]^2$
Where

$$w^i(x) = exp\left(-\frac{||x^{(i)} - x||^2}{2\tau^2}\right)$$
$$b^l(x^{(i)}) = \begin{bmatrix}1 & x_1^{(i)} & x_2^{(i)} & x_3^{(i)} & \cdots & x_n^{(i)}\end{bmatrix}^T$$
$$\theta = \begin{bmatrix}\theta_0 & \theta_1 & x\theta_2 & \theta_3 & \cdots & \theta_n\end{bmatrix}^T$$

Rewriting $\theta^T b^l(x^{(i)})$ as $b^l(x^{(i)})^T\theta$

$$\begin{aligned} E^{LWLR}(\theta) &= \frac{1}{2}\sum_{i=1}^{m} w^{(i)}(x)[y^{(i)} - b^l(x^{(i)})^T\theta]^2 \\ &= \frac{1}{2}\sum_{i=1}^{m} w^{(i)}(x)[b^l(x^{(i)})^T\theta - y^{(i)}]^2 \end{aligned} \tag{1.1}$$

Now Let

$$\mathbf{y} \in \mathbb{R}^{m\times 1} = \begin{bmatrix}y^{(1)} & y^{(2)} & y^{(1)} & y^{(3)} & \cdots & y^{(m)}\end{bmatrix}^T$$

$$B \in \mathbb{R}^{m\times n} = \begin{bmatrix} b^l(x^{(1)})^T \\ b^l(x^{(2)})^T \\ b^l(x^{(3)})^T \\ \vdots \\ b^l(x^{(m)})^T \end{bmatrix}$$

Therefore,

$$B\theta \in \mathbb{R}^{m\times 1} = \begin{bmatrix} b^l(x^{(1)})^T \\ b^l(x^{(2)})^T \\ b^l(x^{(3)})^T \\ \vdots \\ b^l(x^{(m)})^T \end{bmatrix}.\theta$$

Following the definition of $w^i(x)$ above, the weight matrix W is of the form

$$W = \begin{bmatrix} w^1(x) & 0 & 0 & \cdots & 0 \\ 0 & w^2(x) & 0 & \cdots & 0 \\ 0 & 0 & w^3(x) & \cdots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & w^m(x) \end{bmatrix}$$

$W[i,i] = exp(-\frac{||x^{(i)}-x||^2}{2\tau^2})$ and $\forall_{i \neq j} \quad W[i,j] = 0$.

From expression, it can be seen that $W \in \mathbb{R}^{m \times m}$ is a diagonal matrix with the diagonal elements being $w^i(x)$, the distance of x from the *i-th* example.

Using quadratic expressions, Eq. 1.1 can be written as

Solution 1.1

$$E^{LWLR}(\theta) = \frac{1}{2}(B\theta - \boldsymbol{y})^T W(B\theta - y)$$

Where the 0.5 is just a scaling factor.

1.2 q1d

There is underfitting for $\tau > 100$

1.3 q1e

There is no overfitting for the given values of τ.

1.4 q1f

No, the performance of cross validation is better than that of the test error. The superior performance is because because the training set in cross validation changes slightly across the N-folds. That is the distribution of the training examples for the training errors is different from the N-different training examples for the cross validation. Thus, the cross validation is not a good measure of test performance.

2 q2

2.1 q2c

No, the test errors and training errors are not different. The reason is both both variants of gradient ascent converged to the same optimum. That's both of them divided the feature space using the same hyperplane. Hence the resulting similarities in the training and test errors for both methods. This also makes intuitive sense because besides, alpha, nothing changes between the two variants of gradient ascent. And since we are always making progress on alpha, the two methods should converge to approximately the same optima.

2.2 q2d

The method using line-search converged faster than the fix-step vanilla gradient ascent method. The reason is that since the method using line search adapts α at different regions of the optimization curve to speed up optimization. Initially, the algorithm makes huge progress because the parameters are far away from the optimum but as learning takes progresses, line search changes the learning rate to small values which helps it to converge aster. The fix-step method on the other hand has the same momentum throughout the optimization.

3 q3

Answer: The gamer should choose the other unopened door.
$C \in \{\#1, \#2, \#3\}$ Where $p(C = \#1)$ denotes the probability that the car is behind door #1.
$H \in \{\#1, \#2, \#3\}$ Where $p(H = \#2)$ denotes the probability that door #2 is opened by the host.
Before any selection is done, the priors are:

$$p(C = \#1) = p(C = \#2) = p(C = \#3) = \frac{1}{3}$$

cle
Assuming the gamer chooses door 1. ie, $C = \#1$,
The posterior probability the car is indeed behind door 1 is given by

$$\begin{aligned} p(C = \#1|H = \#2 \; or \; H = \#3) &= \frac{p(C = \#1)p(H = \#2|C = \#1)}{p(H)} + \frac{p(C = \#1)p(H = \#3|C = \#1)}{p(H)} \\ &= \frac{p(C = \#1).\{p(H = \#2|C = \#1) + p(H = \#3|C = \#1)\}}{p(H)} \end{aligned}$$

Since the host must opened one of the other two doors not chosen by the gamer, $p(H = 1) = 0$. It therefore follows automatically that

$$p(H = \#2|C = \#1) + p(H = \#3|C = \#1) = 1$$

Hence,

$$p(H) = 1$$

Therefore,

$$p(C = \#1|H = \#2 \cup H = \#3) = \frac{1}{3}.1 = \frac{1}{3} \tag{3.1}$$

From the answer arrived at in Eq. 3.1, it means the gamer still has one-thirds chance of winning the car if he or she sticks with the initial choice of $C = \#1$. Conversely, if the gamer switches choice and assuming $H = \#3$,

Solution 3.1

$$\begin{aligned} p(C = \#2|H = \#3) &= 1 - p(C = \#1|H = \#3) \\ &= 1 - \frac{1}{3} \\ &= \frac{2}{3} \end{aligned}$$

Thus, if the gamer changes choice to door 2,i.e $C = \#2$, he or she has two-thirds chances of winning the car. Hence, the logical decision to take is to choose the other unopened door.

4 q4

4.1 q4a: Deriving the decision rule for binary classifier of real-valued vectors

Proof.
Given a decision rule:

$$y = \begin{cases} 1 & if \ \frac{1}{1+\exp(-a(x))} > \frac{1}{2} \\ 0 & otherwise \end{cases} \tag{4.1}$$

Where

$$a(x) = \log \frac{p(x|y=1)p(y=1)}{p(x|y=0)p(y=0)}$$

From Eq. 4.1, $\frac{1}{1+\exp(-a(x))} > \frac{1}{2}$ iFF $a(x) > 0$

Hence, assuming $a(x) > 0$

$$\begin{aligned} a(x) &> 0 \\ \log \frac{p(x|y=1)p(y=1)}{p(x|y=0)p(y=0)} &> 0 \\ \log(p(x|y=1)p(y=1)) - \log(p(x|y=0)p(y=0)) &> 0 \\ => \log(p(x|y=1)p(y=1)) &> \log(p(x|y=0)p(y=0)) \end{aligned}$$

■

Therefore, because the log is monotonic, it means that whenever the posterior that a given test example is of class 1 is greater than the posterior that it's of class 0, $\frac{1}{1+\exp(-a(x))} > \frac{1}{2}$. Hence the predicted class of that test example should be of class 1. Conversely, $\frac{1}{1+\exp(-a(x))} \leq \frac{1}{2}$ whenever the posterior $p(y=1|x) \leq p(y=0|x)$ which automatically implies the test example belongs to class 0.

4.2 q4b: Proving that the decision boundary of Gaussian Discriminant Analysis is linear

Proof.
Given the maximum likelihood parameters $\{\mu_0, \mu_1, \Sigma\}$ for a given training set containing two classes $y = 1$ and $y = 0$. Where

$$\begin{aligned} \mu_0 \in R^n &= \mathbb{E}(y=0|x) \\ \mu_1 \in R^n &= \mathbb{E}(y=1|x) \\ \Sigma \in R^{n\times n} &= shared \quad covariance.symmetric \quad and \quad positive \quad definite \end{aligned}$$

At the decision boundary, the posterior probabilities of the two class are equal. That is, $p(y = 1|x, \mu_1, \Sigma) = (y = 0|x, \mu_0, \Sigma)$. If $a(x)$ models the decision boundary, such that $a(x) = 0$

$$\begin{aligned}
a(x) &= \log \frac{p(x|y=1;\mu_1,\Sigma)p(y=1)}{p(x|y=0;\mu_0,\Sigma)p(y=0)} \\
&= \log \{x|y=1;\mu_1,\Sigma)p(y=1)\} - \log \{p(x|y=0;\mu_0,\Sigma)p(y=0)\} \\
&= \log p(x|y=1;\mu_1,\Sigma) + \log p(y=1) - \log p(x|y=0;\mu_0,\Sigma) - \log p(y=0) \\
&= \log p(x|y=1;\mu_1,\Sigma) - \log p(x|y=0;\mu_0,\Sigma) + \log \frac{p(y=1)}{p(y=0)}
\end{aligned} \tag{4.2}$$

Since the model is binary linear Gaussian discriminant, the following definitions apply.

Definition 4.1
Let $p(y = 1) = \phi$. Hence, $p(y = 0) = 1 - \phi$

$$\begin{aligned}
p(x|y=1;\mu_1,\Sigma) &\sim \mathcal{N}(x \in R^n;\mu_1,\Sigma) \\
&= \frac{1}{((2\pi)^n|\Sigma|)^{1/2}}.\exp\left\{-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right\} \\
=> \log p(x|y=1;\mu_1,\Sigma) &= \log\left\{\frac{1}{((2\pi)^n|\Sigma|)^{1/2}}.\exp\left\{-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right\}\right\} \\
&= -\frac{1}{2}\log(2\pi)^n|\Sigma| - \frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) \\
If \quad -\frac{1}{2}\log(2\pi)^n|\Sigma| = K & \\
\log p(x|y=1;\mu_1,\Sigma) &= K - \frac{1}{2}(x^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_1 - \mu_1^T\Sigma^{-1}x + \mu_1^T\Sigma^{-1}\mu_1) \\
Since\ \Sigma\ is\ symmetric,\ x^T\Sigma^{-1}\mu_0 = \mu_0^T\Sigma^{-1}x & \\
\therefore \log p(x|y=1;\mu_1,\Sigma) &= K - \frac{1}{2}(x^T\Sigma^{-1}x - 2\mu_1^T\Sigma^{-1}x + \mu_1^T\Sigma^{-1}\mu_1)
\end{aligned} \tag{4.3}$$

Following similar expansion as in Eq. 4.3,

$$\log p(x|y=0;\mu_0,\Sigma) = K - \frac{1}{2}(x^T\Sigma^{-1}x - 2\mu_0^T\Sigma^{-1}x + \mu_0^T\Sigma^{-1}\mu_0) \tag{4.4}$$

Only the means vary between the two class conditionals.

Following the definitions above, Eq. 4.2 becomes

$$\begin{aligned}
a(x) &= K - \frac{1}{2}(x^T\Sigma^{-1}x - 2\mu_1^T\Sigma^{-1}x + \mu_1^T\Sigma^{-1}\mu_1) - (K - \frac{1}{2}(x^T\Sigma^{-1}x - 2\mu_0^T\Sigma^{-1}x + \mu_0^T\Sigma^{-1}\mu_0)) + \log\frac{\phi}{1-\phi} \\
&= -\frac{1}{2}(x^T\Sigma^{-1}x - 2\mu_1^T\Sigma^{-1}x + \mu_1^T\Sigma^{-1}\mu_1) + \frac{1}{2}(x^T\Sigma^{-1}x - 2\mu_0^T\Sigma^{-1}x + \mu_0^T\Sigma^{-1}\mu_0) + \log\frac{\phi}{1-\phi} \\
&= \mu_1^T\Sigma^{-1}x - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}x + \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 + \log\frac{\phi}{1-\phi} \\
&= \mu_1^T\Sigma^{-1}x - \mu_0^T\Sigma^{-1}x - \frac{1}{2}(\mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0) + \log\frac{\phi}{1-\phi} \\
&= (\mu_1^T - \mu_0^T)\Sigma^{-1}x - -\frac{1}{2}(\mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0) + \log\frac{\phi}{1-\phi} \\
&= (\mu_1 - \mu_0)^T\Sigma^{-1}x - -\frac{1}{2}(\mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0) + \log\frac{\phi}{1-\phi}
\end{aligned} \tag{4.5}$$

From Eq. 4.5, let

$$\begin{aligned}
\theta^T &= (\mu_1 - \mu_0)^T\Sigma^{-1} \qquad And \\
b &= -\frac{1}{2}\left(\mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0\right) + \log\frac{\phi}{1-\phi}
\end{aligned}$$

$$\theta^T \in R^{1\times n} \quad and \quad b \in R^{1\times 1}$$

Hence, Eq 4.5 can be written as

$$a(x) = \theta^T x + b \qquad (\textbf{proved}). \tag{4.6}$$

■

Thus, it's been shown that the decision boundary of a generative Linear Gaussian Discriminant classifier is linear in the input space given the maximum likelihood parameters μ_0, μ_1, Σ.