R Notebook

Q1

1. **ESL 4.2: Similarity of LDA and linear regression for two classes**
   In this problem you will show that for two classes, linear regression leads to the same discriminating direction as LDA, but not to the exact same classification rule in general.
   The derivations for this problem are rather lengthy. Consider part (b) (finding the linear regression direction) to be extra credit. If you fail to prove one step, try to comment on its geometric interpretation instead, and move to the next step.

WLOG assume class 1 is coded as $-N/N_1$ and class 2 as N/N_2, also assume that each paramater is estimated by the unbiased estimator
we know from calss that the LDA border is:

$$x^t\Sigma^{-1}(\mu_2-\mu_1) > 0.5\mu_2^t\Sigma^{-1}\mu_2 - 0.5\mu_1^t\Sigma^{-1}\mu_1 + ln(\frac{N_1}{N}) - ln(\frac{N_2}{N})$$

lets take a closer look at the estimator for β_{OLS} (with intercept)

$$X^tX\begin{bmatrix}\beta_0\\ \beta\end{bmatrix} = X^TY$$

$$X^TY = \begin{bmatrix}1 & 1 & 1 & \dots & 1\\ x_1 & x_2 & x_3 & \dots & x_n\end{bmatrix}\begin{bmatrix}-N/N_1\\ \vdots\\ -N/N_1\\ N/N_2\\ \vdots\\ N/N_2\end{bmatrix} = \begin{bmatrix}N_1(-N/N_1)+N_2(N/N_2)\\ (-N/N_1)\sum_{i=1}^{N_1}x_i + (N/N_2)\sum_{i=N_1+1}^{N_2}x_i\end{bmatrix} = \begin{bmatrix}0\\ N(\mu_2-\mu_1)\end{bmatrix}$$

$$X^TX = \begin{bmatrix}1 & 1 & 1 & \dots & 1\\ x_1 & x_2 & x_3 & \dots & x_n\end{bmatrix}\begin{bmatrix}1 & x_1^t\\ \vdots & \vdots\\ 1 & x_n^T\end{bmatrix} = \begin{bmatrix}N & \sum_{i=1}^N x_i\\ \sum_{i=1}^N x_i & \sum_{i=1}^N x_ix_i^t\end{bmatrix} = \begin{bmatrix}N & N_1\mu_1+N_2\mu_2\\ N_1\mu_1+N_2\mu_2 & \sum_{i=1}^N x_ix_i^t\end{bmatrix}$$

$$X^tX\begin{bmatrix}\beta_0\\ \beta\end{bmatrix} = \begin{bmatrix}N\beta_0 + (N_1\mu_1^t+N_2\mu_2^t)\beta\\ (N_1\mu_1+N_2\mu_2)\beta_0 + \sum_{i=1}^N x_ix_i^t\beta\end{bmatrix} = \begin{bmatrix}0\\ N(\mu_2-\mu_1)\end{bmatrix}$$

$$N\beta_0 + (N_1\mu_1^t+N_2\mu_2^t)\beta = 0 \Rightarrow \beta_0 = -\frac{(N_1\mu_1^t+N_2\mu_2^t)}{N}\beta$$

plug it back in:

$$(N_1\mu_1+N_2\mu_2)\beta_0 + \sum_{i=1}^N x_ix_i^t\beta = [(N_1\mu_1+N_2\mu_2)\frac{-(N_1\mu_1^t+N_2\mu_2^t)}{N} + \sum_{i=1}^N x_ix_i^t]\beta = N(\mu_2-\mu_1)$$

$$\Sigma = \frac{1}{N-2}[\sum_{i=1}^{N_1}(x_i-\mu_1)(x_i-\mu_1)^T + \sum_{i=N_1+1}^{N_2}(x_i-\mu_2)(x_i-\mu_2)^T] =$$

$$= \frac{1}{N-2}[\sum_{i=1}^{N_1}x_ix_i^t - \sum_{i=1}^{N_1}x_i\mu_1^t - \sum_{i=1}^{N_1}\mu_1x_i^t + N_1\mu_1\mu_1^t + \sum_{i=N_1+1}^{N_2}x_ix_i^t - \sum_{i=N_1+1}^{N_2}x_i\mu_2^t - \sum_{i=N_1+1}^{N_2}\mu_2x_i^t + N_2\mu_2\mu_2^t] = \frac{1}{N-2}[\sum_{i=1}^{N}x_ix_i^t - N_1\mu_1\mu_1^t - N$$

we get: $\sum_{i=1}^N x_ix_i^t = (N-2)\Sigma + N_1\mu_1\mu_1^t + N_2\mu_2\mu_2^t$

plug it back in and we get:

$$[(N_1\mu_1+N_2\mu_2)\frac{-(N_1\mu_1^t+N_2\mu_2^t)}{N} + (N-2)\Sigma + N_1\mu_1\mu_1^t + N_2\mu_2\mu_2^t]\beta = N(\mu_2-\mu_1) \Rightarrow \beta \propto \Sigma^{-1}(\mu_2-\mu_1)$$

thus the lines are in the same directions but are not identical beacuse $\beta \neq \Sigma^{-1}(\mu_2-\mu_1)$

Q2

2. **Short intuition problems**
Choose and explain briefly. If you need additional assumptions to reach your conclusion, specify them.

(a) What is not an advantage of using logistic loss over using squared error loss with 0-1 coding for 2-class classification?

i. That the expected prediction error is minimized by correctly predicting $P(Y|X)$.
ii. That it has a natural probabilistic generalization to $K > 2$ classes.
iii. That its predictions are always legal probabilities in the range $(0, 1)$.

(b) In the generative 2-class classification models LDA and QDA, what type of distribution does $P(Y|X = x)$ have?

i. Unknown
ii. Gaussian
iii. Bernoulli

(c) We mentioned in class that Naive Bayes assumes $P(\mathbf{x}|Y = g) = \Pi_{j=1}^{p} P_j(x_j|Y = g)$. In what situation would you expect this simplifying assumption to be most useful?

i. Small number of predictors, not highly correlated.
ii. Small number of predictors, highly correlated between them.
iii. Large number of predictors, not highly correlated.
iv. Large number of predictors, many highly correlated between them.

a

the answer is i, by predicting correctly P(Y|X) we also minimize the EPE in the linear regression case.

b

the answer is iii - Bernoulli
in the 2 class classification {1,-1}: $P(Y = 1|X = x) = 1 - P(Y = -1|X = x) = p < 1 \Rightarrow$ the sidtribution is Bernoulli

c

the answer is i - Small number of predictors, not highly correlated.
if the observations x are highly correlated then $P(x|Y = g) \neq \Pi_{j=1}^{p} Pj(xj|Y = g)$.
if the number of predictors is high this expression will go quicly to zero thus the probabilty will be very low for all x.

Q3

3. **Equivalence of selecting "reference class" in multinomial logistic regression**
In class we defined the logistic model as:

$$\begin{aligned} \log\left(\frac{P(G=1|X)}{P(G=K|X)}\right) &= X^T\beta_1 \\ &\vdots \\ \log\left(\frac{P(G=K-1|X)}{P(G=K|X)}\right) &= X^T\beta_{K-1}, \end{aligned}$$

with resulting probabilities:

$$\begin{aligned} P(G=k|X) &= \frac{\exp\{X^T\beta_k\}}{1+\sum_{l<K}\exp\{X^T\beta_l\}}, \quad k<K \\ P(G=K|X) &= \frac{1}{1+\sum_{l<K}\exp\{X^T\beta_l\}}. \end{aligned}$$

Show that if we choose a different class in the denominator, we can obtain the same set of probabilities by a different set of linear models (i.e., values of β). Hence the two representations are equivalent in the probabilities they yield.

wlog lets show that we can switch the denominator to calss 1

$$\text{difine } x^t\tilde{\beta}_i = ln(\frac{P(G=i|x)}{P(G=1|x)}) \quad \forall\, k > i > 2$$

$$ln(P(G=1|x)) - ln(P(G=K|x)) = x^t\beta_1$$

$$ln(P(G=K|x)) - ln(P(G=1|x)) = x^t\tilde{\beta}_K$$

$$\text{summing the 2 aboice equation yields: } x^t\beta_1 = -x^t\tilde{\beta}_K$$

$$x^t\tilde{\beta}_i = ln(\frac{P(G=i|x)}{P(G=1|x)}) = ln(P(G=i|x)) - ln(P(G=1|x)) + ln(P(G=K|x)) - ln(P(G=K|x)) =$$

$$= ln(\frac{P(G=i|x)}{P(G=K|x)}) + -x^t\beta_1 = x^t(\beta_i - \beta_1)$$

$$\text{thus we get that } \tilde{\beta}_i = \beta_i - \beta_1 \;\&\; \tilde{\beta}_K = -\beta_1$$

this is the new linear model, we are left with shoing that the probabilities are the same

$$P(G=k|x)[\text{under class K as refrence}] = \frac{e^{x^t\beta_k}}{1+\sum_{l<K} e^{x^t\beta_l}} = \frac{e^{x^t\beta_k}}{1+\sum_{l<K} e^{x^t\beta_l}}\frac{e^{-x^t\beta_1}}{e^{-x^t\beta_1}} = \frac{e^{x^t\tilde{\beta}_k}}{e^{-x^t\beta_1}+\sum_{l<K} e^{x^t\tilde{\beta}_l}} =_* \frac{e^{x^t\tilde{\beta}_k}}{1+\sum_{1<l} e^{x^t\tilde{\beta}_l}} =$$

$$= P(G=k|x)[\text{under class 1 as refrence}]$$

$$*: e^{-x^t\beta_1} + \sum_{l<K} e^{x^t\tilde{\beta}_l} = e^{-x^t\beta_1} + \sum_{1<l<K} e^{x^t\tilde{\beta}_l} + e^{x^t\tilde{\beta}_1} = e^{x^t\tilde{\beta}_K} + \sum_{1<l<K} e^{x^t\tilde{\beta}_l} + e^{x^t 0} = 1 + \sum_{1<l} e^{x^t\tilde{\beta}_l}$$

Q4

4. **Separability and optimal separators**
   ESL 4.5: Show that the solution of logistic regression is undefined if the data are separable.

lets show it to the 2 dimentional case (classes 0,1), the p>2 dimentional case is equivelent (with p-1 equivilent equations)

$$\hat{\beta} = argmax \sum_{i=1}^{n} I(y_i = 1)(x_i^t\beta) - log(1 + exp(x_i^t\beta))$$

$$\text{denote } L = x^t\hat{\beta}y - \sum_{i=1}^{n} log(1 + exp(x_i^t\hat{\beta}))$$

$$\text{in the separable case: } \sum_{i=1}^{n} I(y_i = 1)(x_i^t\beta) = x^t\beta y > 0$$

$$\text{we will classafy a point as 1 if: } x^t\hat{\beta} > 0$$

$$\text{and for } m > 0 : x^t\hat{\beta} > 0 \iff x^t m\hat{\beta} > 0$$

$$mx^t\hat{\beta}y - \sum_{i=1}^{n} log(1 + exp(x_i^t\hat{\beta})e^m) > mx^t\hat{\beta}y - \sum_{i=1}^{n} log([1 + exp(x_i^t\hat{\beta})]e^m) = m(x^t\hat{\beta}y - n) - \sum_{i=1}^{n} log(1 + exp(x_i^t\hat{\beta})) >$$

$$> m(x^t\hat{\beta}y - \sum_{i=1}^{n} log(1 + exp(x_i^t\hat{\beta})) - n) = m(L - n) > L \iff m > \frac{L}{L-n}$$

$$\text{so if we take } m = \frac{L+1}{L-n} \text{ and plug } m\hat{\beta} \text{ into the loss function we get a higer value thus } \hat{\beta} \text{ is not argmax}$$

$$(*)\; x^t\hat{\beta} < log(1 + exp(x_i^t\hat{\beta})) \Rightarrow L < 0$$

Q5

5. **(* A real challenge[1])**
   In the separable case, consider adding a small amount of ridge-type regularization to the likelihood:

$$\hat{\beta}(\lambda) = \arg\min_{\beta} -l(\beta; X, \mathbf{y}) + \lambda \sum_j \beta_j^2$$

where $l(\beta; X, \mathbf{y})$ is the standard logistic log likelihood.
Show that $\hat{\beta}(\lambda)/\|\hat{\beta}(\lambda)\|_2$ converges to the hard-margin support vector classification solution (margin maximizing hyper-plane) as $\lambda \to 0$.
Hint:You may find the equivalent formulation of SVM in equation (4.48) of ESL (Second Edition) useful.