Statistical Learning, Fall 2022

# Homework exercise 3

Due date: 20 December in class

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.

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.

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 $\beta$). Hence the two representations are equivalent in the probabilities they yield.

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

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.

[1] +50 points extra credit for original solution; +20 points for finding a solution in the literature and explaining it clearly; +5 for finding and citing it only