Statistical Learning, Spring 17

Homework exercise 3

Due date: 29 May 2015

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 β). 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 support vector machine solution (margin maximizing hyperplane) as $\lambda \to 0$.
Hint:You may find the equivalent formulation of SVM in equation (4.44) of ESL useful (equation (4.48) in the book's second Edition).

6. **Questions on class presentations from 19 May**

   (a) **Statistical vs. contextual model evaluation**
   Consider the two evaluation approaches discussed in the wallet estimation case study. In slides 20–21 we used publicly available datasets and quantile loss on holdout (validation) sets to compare performance of various approaches. In slides 29–30 we compared various models' wallet estimates to experts' "validated opportunities".

      i. Explain briefly why both evaluation approaches are necessary for comparison and validation of modeling approaches.

      ii. Which approach would be more appropriate for publication in an applied statistics or machine learning journal? In what kind of forum would the other approach be likely to be positively accepted?

   For more details about the evaluation setup and results, you can look at the paper (available from my home page):
   C. Perlich, S. Rosset, R. Lawrence, B. Zadrozny. *High Quantile Modeling for Customer Wallet Estimation with Other Applications.*

   (b) **Yehuda Koren's presentation on Netflix $1M competition:**

      i. Consider the "collaborative filtering" model on slide 26. Assume we want to use this method on Netflix data. Explain why either an intercept or overall standardization of the r_{ui}'s is required rather than using the raw ratings in $\{1, 2, 3, 4, 5\}$ (you may find it useful to think about the vectors p, q, as embedded points as on slide 19).

[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

ii. Slides 37, 38 give two views of temporal behavior. How can 37 can be used to explain 38?

* Extra credit: Some claims have been made that the results on Slide 38 make no sense, because the grade for each movie tends to decrease over time, rather than increase. Assuming that this is true, an overall pattern like in slide 38 can still occur, despite the apparent contradiction. Can you suggest an explanation?

* Extra credit: On slide 42 we see better prediction for "active" users. This can conceivably be due to two distinct reasons: the nature of the users ("low bias") or the use of more information to better model active users ("low variance"). Explain briefly the two options and suggest how the two can be differentiated with a simple modeling exercise.