Statistical Machine Learning GU4241/GR5241
Spring 2017
https://courseworks.columbia.edu/

Linxi Liu
ll3098@columbia.edu
Shuaiwen Wang
sw2853@columbia.edu
Haolei Weng
hw2375@columbia.edu

Homework 2

Due: Tuesday, Feb. 21st, 2017

Homework submission: We will collect your homework **at the beginning of class** on the due date. You also need to submit **your code** for Problem 3 through Canvas. If you cannot attend class that day, you can leave your solution in the homework dropbox 904 at the Department of Statistics, 9th floor SSW, at any time before then.

Problem 1 (Bayes-Optimal Classifier, 10 points)

Consider a classification problem with K classes and with observations in $\mathbb{R}^d$. Now suppose we have access to the true joint density $p(\mathbf{x}, y)$ of the data $\mathbf{x}$ and the labels y. From $p(\mathbf{x}, y)$ we can derive the conditional probability $P(y|\mathbf{x})$, that is, the posterior probability of class y given observation $\mathbf{x}$.

In the lecture, we have introduced a classifier f_0 based on p, defined as

$$f_0(\mathbf{x}) := \arg\max_{y \in [K]} P(y|\mathbf{x}) \ ,$$

the *Bayes-optimal classifier*.

Homework question: Show that the Bayes-optimal classifier is the classifier which minimizes the probability of error, under all classifiers in the hypothesis class

$$\mathcal{H} := \{ f \colon \mathbb{R}^d \to [K] \,|\, f \text{ integrable } \} \ .$$

(If you are not familiar with the notion of an integrable function, just think of this as the set of all functions from $\mathbb{R}^d$ to the set $[K]$ of class labels.)

Hints:
- The probability of error is precisely the risk under zero-one loss.
- You can greatly simplify the problem by decomposing the risk $R(f)$ into conditional risks $R(f|\mathbf{x})$:

$$R(f|\mathbf{x}) := \sum_{y \in [K]} L^{0\text{-}1}(y, f(\mathbf{x})) P(y|\mathbf{x}) \qquad \text{and hence} \qquad R(f) = \int_{\mathbb{R}^d} R(f|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \ .$$

  If you can show that f_0 minimizes $R(f|\mathbf{x})$ for every $\mathbf{x} \in \mathbb{R}^d$, the result for $R(f)$ follows by monotonicity of the integral.

Problem 2 (Non-linear Decision Boudary, ISL 9.2, 10 points)

We have seen that in $p = 2$ dimensions, a linear decision boundary takes the form $\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$. We now investigate a non-linear decision boundary.

(a) Sketh the curve

$$(1 + X_1)^2 + (2 - X_2)^2 = 4.$$

(b) On your sketch, indicate the set of points for which

$$(1 + X_1)^2 + (2 - X_2)^2 > 4,$$

as well as the set of points for which

$$(1 + X_1)^2 + (2 - X_2)^2 \leq 4.$$

(c) Suppose that a classifier assigns an observation to the blue class if

$$(1 + X_1)^2 + (2 - X_2)^2 > 4,$$

and to the red class otherwise. To what class is the observation $(0, 0)$ classified? What about $(-1, 1)$, $(2, 2)$ or $(3, 8)$?

(d) Argue that while the decision boundary in (c) is not linear in terms of X_1 and X_2, it is linear in terms of X_1, X_1^2, X_2 and X_2^2.

Problem 3 (LDA and Logistic Regression, 20 points)

The zipcode data are high dimensional, and hence linear discriminant analysis suffers from high variance. Using the training and test data for the 3s, 5s, and 8s, compare the following procedures:

1. LDA on the original 256 dimensional space.
2. LDA on the leading 49 principle components of the features.
3. LDA when you *filter* the data as follows. Each non-overlapping 2×2 pixel block is replaced by its average.
4. Multiple linear logistic regression using the same filtered data as in the previous question. [Use the `multinomial` family in the R package `glmnet`; use the solution at the end of the path].

Homework Problems. Compare the procedures with respect to training and test misclassification error. You need to report both training and test misclassification error in your submission.