Homework 7

Advanced Methods for Data Analysis (36-402/36-608)

Due Tues April 8, 2014 at 11:59pm

Instructions: this homework has no programming part, so you only have to submit solutions to the following exercises. As usual, you must submit a PDF; any other file type will not be accepted.

1 Principal component analysis problems

1.1 Sample means

(a) Let $a \in \mathbb{R}^n$ be a vector. Show that a has sample mean (i.e., the components of a have sample mean) equal to

$$\bar{a} = \frac{1}{n} \mathbb{1}^T a,$$

where $\mathbb{1}$ is the $n \times 1$ vector of all 1s.

(b) Suppose that $X \in \mathbb{R}^{n \times p}$ is a matrix whose columns are centered, i.e., have sample mean zero. Show that

$$\mathbb{1}^T X = 0,$$

where in the above, the right-hand side denotes the $1 \times p$ vector of all 0s.

(c) Now let $v \in \mathbb{R}^p$ be an arbitrary vector, and $X \in \mathbb{R}^{n \times p}$ be a matrix as above whose columns are centered. Show that the vector Xv has sample mean zero. Hint: use parts (a) and (b).

1.2 Orthogonality and directions

(a) Suppose that $v_1, \ldots v_k \in \mathbb{R}^p$ are orthogonal, meaning that $v_i^T v_j = 0$ whenever $i \neq j$. Show that $v_1, \ldots v_k$ are linearly independent vectors.

(b) Use part (a) to argue that there cannot exist more than p orthogonal vectors in $\mathbb{R}^p$.

(c) Use part (b) to argue that there cannot exist more than p principal component directions for a given data matrix $X \in \mathbb{R}^{n \times p}$.

1.3 Total sample variance

(a) Suppose that $X \in \mathbb{R}^{n \times p}$ is a data matrix with centered columns. Note that the sample variance of the data points (rows) in X, along the jth dimension, is given by

$$\frac{1}{n} \sum_{i=1}^{n} X_{ij}^2,$$

for $j = 1, \dots p$. Define the *total sample variance* of X to be the sum of the sample variances along each of the p dimensions, i.e.,

$$\mathrm{TotVar}(X) = \frac{1}{n} \sum_{j=1}^{p} \sum_{i=1}^{n} X_{ij}^2.$$

Show that the total sample variance can be written as $\mathrm{TotVar}(X) = \mathrm{tr}(\frac{1}{n} X^T X)$, where recall that $\mathrm{tr}(A)$ denotes the trace of a matrix A, i.e., the sum of its diagonal elements.

(b) Let X have singular value decomposition $X = UDV^T$, where $U \in \mathbb{R}^{n \times p}$ has orthonormal columns, $D \in \mathbb{R}^{p \times p}$ is diagonal with diagonal elements $d_1 \geq \ldots \geq d_p \geq 0$, and $V \in \mathbb{R}^{p \times p}$ has orthonormal columns. Prove that the total sample variance of X is

$$\mathrm{TotVar}(X) = \frac{1}{n} \sum_{j=1}^{p} d_j^2.$$

Hint: start with the result from part (a). Also, use the fact that you can commute the product of matrices under the trace operation, i.e., $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.

2 General review problems

2.1 Orthonormal linear regression

(a) Suppose that we are given an outcome vector $y \in \mathbb{R}^n$ and predictor matrix $X \in \mathbb{R}^{n \times p}$, where X has orthonormal predictors (i.e., orthonormal columns). Prove that the linear regression coefficients of y on X are given simply by taking the inner product of each predictor with y (i.e., each column with y).

(b) Write the columns of X as $X_1, \dots X_p \in \mathbb{R}^n$. Let $\hat{\beta}$ denote the coefficient vector from regressing y on X. Use part (a) to show that $\hat{\beta}_j$ is the same as the coefficient from regressing y on X_j, the output of a *univariate linear regression*, for each $j = 1, \dots p$. Note that here we mean univariate linear regression without intercept.

(c) What does this tell you about dropping variables, say, when looking at p-values, from a linear regression of y on orthonormal predictors?

2.2 Variance estimation in nonparametric regression

Consider the nonparametric model

$$y_i = r(x_i) + \epsilon_i, \quad i = 1, \dots n,$$

where $x_1, \dots x_n$ are considered fixed, and $\epsilon_1, \dots \epsilon_n$ are i.i.d. with mean 0 and variance σ^2.

(a) Suppose that we observe an additional copy of this data set

$$y_i' = r(x_i) + \epsilon_i', \quad i = 1, \dots n,$$

where now all errors $\epsilon_1, \dots \epsilon_n, \epsilon_1', \dots \epsilon_n'$ are i.i.d. with mean 0 and variance σ^2. Consider the following variance estimator:

$$T = \frac{1}{2n} \sum_{i=1}^{n} (y_i - y_i')^2.$$

Prove that $\mathbb{E}(T) = \sigma^2$. Hint: consider just one term at a time in the sum. Now, add and subtract a key quantity in each term.

(b) Explain in words (but still, concretely) why you would prefer the estimator T in part (a) over the simpler estimator $\frac{1}{2}(y_1 - y_1')^2$.

(c) Now suppose that we didn't have an extra copy of our data set. One alternative idea, assuming that $x_1 < x_2 < \ldots < x_n$, is to use the estimator

$$U = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (y_i - y_{i+1})^2.$$

Derive an expression for $\mathbb{E}(U)$, simplifying as much as possible.

(d) Under what circumstances would you think U is a good estimator, i.e., would $\mathbb{E}(U)$ be close to σ^2?

2.3 True or false

You only have to answer true or false for each of the following questions. (As practice, you can try answering these without consulting your notes.)

1. The kernel smoothing estimate with infinite bandwidth is simply a linear regression fit to the data samples.
2. The smoothing spline estimate with infinite smoothing parameter is simply a linear regression fit to the data samples.
3. Lower training error generally means a better method.
4. If we run K-fold cross-validation for some regression method, writing $\text{CV}_1, \ldots \text{CV}_K$ to denote the errors from the K folds, then the cross-validation error estimate

   $$\text{CVErr} = \frac{1}{K} \sum_{k=1}^{K} \text{CV}_k$$

   is exactly an unbiased estimate of expected test error.
5. An appropriate estimate for the standard deviation of CVErr is given by the sample standard deviation of $\text{CV}_1, \ldots \text{CV}_K$.
6. If two methods A and B have the same degrees of freedom, but method A has a higher training error than method B, then we have good reason to believe that method A will also have a higher test error than B.
7. Additive models typically suffer from poor variance, but have very low bias.
8. The decision boundary for a logistic regression classifier is defined by the set of all $x \in \mathbb{R}^p$ for which the predicted probability of $Y = 1$, conditional on $X = x$, is equal to $1/2$.
9. Generalized linear models are designed for cases in which the outcome Y isn't exactly a linear function of the predictors X, but rather, Y is a linear function of some transformation of the predictors $g(X)$.
10. If Y is distributed according to an exponential family with natural parameter θ and dispersion parameter ϕ, then the mean $\mu = \mathbb{E}(Y)$ can depend on both θ, ϕ.