10.R Review Questions

10.R.1

1/1 point (graded)

Suppose we want to fit a linear regression, but the number of variables is much larger than the number of observations. In some cases, we may improve the fit by reducing the dimension of the features before.

In this problem, we use a data set with n = 300 and p = 200, so we have more observations than variables, but not by much. Load the data x, y, x.test, and y.test from 10.R.RData.

First, concatenate x and x.test using the rbind functions and perform a principal components analysis on the concatenated data frame (use the "scale=TRUE" option). To within 10% relative error, what proportion of the variance is explained by the first five principal components?

0.34986 ✔ **Answer:** 0.34986

0.34986

Explanation

use "vars = prcomp(rbind(x,x.test),scale=TRUE)$sdev^2"

Submit

Answers are displayed within the problem

10.R.2

1/1 point (graded)

The previous answer suggests that a relatively small number of "latent variables" account for a substantial fraction of the features' variability. We might believe

that these latent variables are more important than linear combinations of the features that have low variance.

We can try forgetting about the raw features and using the first five principal components (computed on rbind(x,x.test)) instead as low-dimensional derived features. What is the mean-squared test error if we regress y on the first five principal components, and use the resulting model to predict y.test?

0.9923 ✔ **Answer:** 0.9923

0.9923

Explanation
In the actual data generating model for this example, the features may be noisy proxies for a few latent variables that actually drive the response. This is not an uncommon situation when we have high-dimensional data.

Submit

Answers are displayed within the problem

10.R.3

1/1 point (graded)
Now, try an OLS linear regression of y on the matrix x. What is the mean squared predition error if we use the fitted model to predict y.test from x.test?

3.657 ✔ **Answer:** 3.657

3.657

Explanation
The mean squared error is worse because of the variance involved in fitting a very high dimensional model. As it turned out here, the large-variance directions of x turned out to be the important ones for predicting y. Note that this need not always be the case, but it often is.

Submit

Answers are displayed within the problem