Stat 432 Homework 4

Assigned: Feb 17, 2019; Due: Feb 22, 2019

Question 1 (more on PCA)

[3 points] In our lectures, we demonstrated that PCA has a close connection with the singular value decomposition (`svd`). Let us verify this connection. Take the first four columns of the `iris` data.

Based on our understanding of the connection between PCA and SVD, obtain the following objects using the `svd()` function. Note that you cannot use any built-in function that performs PCA directly, however, you can use them to check the answer. You can find those answers in the `pca` example document.

- Plot the variances of the principal components in a decreasing order.
- Obtain the first two principal components, and plot them on a figure, color the points with true species.
- Print the rotation matrix.

Question 2 (k-NN for classification)

[4 points] Consider again the zip code digits data. And we will use the Eucleadian distance. We want to predict the digit of the 4th observation in the testing dataset.

```
library(ElemStatLearn)
train.x = zip.train[, -1]
train.y = as.factor(zip.train[, 1])
test.x.one = zip.test[4, -1]
```

Do the following steps. **Note** that you cannot use any built-in kNN function for this entire question. For step 1), you **cannot** use any for-loops. As a hint, you may consider using `sweep` and `rowSums`, while other functions can also get the job done.

1. Using covariates `test.x.one` find the indices of all 15 nearest neighbors in the training data.
2. Find the most frequent digit among these 15 observations. Is this the true digit of this testing data?
3. How about changing the value of k? Can we get a correct prediction?

Apply steps 1 and 2 to the first 100 observations in the testing data, with k ranging from 1 to 20. Which k seems to perform the best? Use evidence to support your answer.

Question 3 (Cross-validation using the `caret` package)

[3 points] The `caret` package is frequently used for cross-validation and choosing tuning parameters of machine learning models. In addition to what we covered in class, read two functions `train()` and `trainControl()` from the `caret` package documentation (link). You may also google any examples to help you. Use these functions to perform a cross-validation of $k$nn on the zip code digit training data (not `zip.test`). You need to use the following configurations when doing this cross-validation.

- possible values of k: all values from 1 to 10
- resampling method: cross-validation
- number of folds: 3

After selecting the best k, use this k to fit the model using the training data and predict on the testing data `zip.test`. Report the prediction confusing matrix.