Pattern Classification and Recognition
ECE 681

Spring 2016
Homework #2: Cross-Validation and Feature Selection

Due: 4:30PM, Thursday, March 3, 2016[1]

This homework assignment is worth **220 points**. (220/220 = 97%)
Up to **10 points** extra credit may be earned by going above and beyond the given problem statements (*e.g.*, performing additional analyses, or providing additional insightful interpration of the results).

Your homework is not considered submitted until all three components (hard-copy, Matlab `.m` code, and Blind Test Results Matlab `.mat` file) have been submitted.

Submit a **hard-copy** with your plots and commentary/interpretations to the homework box in Teer.
Submit your **Matlab `.m` code** as an Attachment to the Assignment in Sakai.
Submit your **Blind Test Results** in a Matlab `.mat` file as an Attachment to the Assignment in Sakai.

Cross-Validation

Implement your own cross-validation function.

Consider a KNN classifier, that uses the L_2-norm as the distance metric, trained (developed) using the following training data:
(this is the same KNN classifier that was considered for **KNN Classification** in Homework #1)

> H_0 features are 50 samples drawn from a 2-dimensional normal (Gaussian) distribution with mean $\mu = [0\ 0]'$ and covariance matrix $C = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ (the identity matrix).
> H_1 features are 50 samples drawn from a 2-dimensional normal (Gaussian) distribution with mean $\mu = [2\ 3]'$ and covariance matrix $C = \left[\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix}\right]$.

Apply 10-folds cross-validation to estimate the KNN classifier's performance when applied to the training data for each of the following for each of the following choices of k: 1, 3, 5, 7, 11, 15, and 19.

(10) 1. Plot cross-validated ROC curves for the KNN classifier when applied to the training data for each choice of k. (You may plot all the ROCs on a single set of axes, provided that you use different line types and include a legend.)

[1] I (Dr. Tantum) will collect homework from the homework box after 4:30PM, allowing for an appropriate grace period.
DO NOT submit late work to the locked homework box in Teer.
Late work is to be submitted to me (Dr. Tantum) in person, to my mailbox in Hudson 130, or slid under my office door (Hudson 114).

> Submitted by 4:30PM, Friday, March 4, 2016 = 1 day late.
> Submitted by 4:30PM, Monday, March 7, 2016 = 2 days late.
> Submitted by 4:30PM, Tuesday, March 8, 2016 = 3 days late.
> Late submissions not accepted after 4:30PM, Tuesday, March 8, 2016.

Work submitted in person to me (Dr. Tantum), to my mailbox in Hudson 130, or slid under my office door (Hudson 114) after the submission deadline but prior to my collecting homework from the homework box will be treated as if it were submitted on time.
Work submitted to the homework box after I have collected homework from the box will receive zero credit.

(10) 2. Plot the area under the ROC curve (AUC) for the mean (cross-validated) ROCs as a function of k when the KNN classifier is applied to the training data.

(10) 3. How do the cross-validated AUC performance estimates as a function of k compare to the performance estimates obtained with incestuous training/testing (*i.e.*, the first part of the results for Question 4 in the **KNN Classification** section of Homework #1, in which KNN performance was evaluated by applying it to the training data)?

(10) 4. How do the cross-validated AUC performance estimates as a function of k compare to the performance estimates obtained with previously unseen testing data (*i.e.*, the second part of the results for Question 4 in the **KNN Classification** section of Homework #1, in which KNN performance was evaluated by applying it to the separate testing data)?

(20) 5. If you were to repeat the blind test completed for Homework #1, (I am not asking you to repeat the blind test!), what value of k would you select for your KNN classifier now, and why would you select it?

Comment on the trade-offs you considered when choosing k, and justify your choice of k.

(30) 6. Submit the Matlab code that produced the above results as an Attachment to the Assignment in Sakai. (We should be able to run this code to replicate your results.)

Feature Selection

Implement your own feature selection function (or functions), such that you can select features by a scalar search, forward sequential search, or exhaustive search.

Generate training data according to the following distributions:
(HINT: The Matlab function `mvnrnd` may be helpful.)

H_0 features are 250 samples drawn from a normal (Gaussian) distribution with mean vector $\mu = [0\ 0\ 0\ 0]^T$ and covariance matrix Σ given below.
H_1 features are 250 samples drawn from a normal (Gaussian) distribution with mean vector $\mu = [0\ 2\ 3\ 3]^T$ and covariance matrix Σ given below.
The covariance matrix for both H_0 and H_1 features is given by

$$\Sigma = \begin{bmatrix} 0.5 & 0 & 0 & 0.3 \\ 0 & 1.0 & 0 & 0.7 \\ 0 & 0 & 1.0 & 0.2 \\ 0.3 & 0.7 & 0.2 & 1.5 \end{bmatrix}$$

Using a KNN classifier with $k = 9$, 10-Folds cross-validation, and AUC for the mean (cross-validated) ROC curve as the performance metric:

(10) 1. Determine the best feature set for this data using a scalar forward search. (That is, add features one at a time where the feature that is added is the remaining feature that is the best by itself.)

(10) 2. Explain why the feature set identified through a scalar forward search is selected as the best feature set. (HINT: Plot AUC for the feature sets considered.)

(10) 3. Determine the best feature set for this data using an exhaustive search. (That is, evaluate performance for all 15 possible feature sets.)

(10) 4. Explain why the feature set identified through an exhaustive search is selected as the best feature set. (HINT: Plot AUC for the feature sets considered.)

(10) 5. Determine the best feature set for this data using a sequential forward search. (That is, find the best single feature, then find the best set of two features that includes the first, then the best set of three features that includes the first two, and so on.)

(10) 6. Explain why the feature set identified through a sequential forward search is selected as the best feature set. (HINT: Plot AUC for the feature sets considered.)

(30) 7. Submit the Matlab code that produced the above results as an Attachment to the Assignment in Sakai. (We should be able to run this code to replicate your results.)

Blind Test

Apply a KNN classifier developed (trained) using training data as specified in **Feature Selection** to generate decision statistics for blind test data.

(20) 1. Select a feature set based on the performance evaluation completed for **Feature Selection**. (Feel free to consider additional values for k beyond that specified for **Feature Selection** if you think if would be helpful or beneficial.)

(20) 2. Generate decision statistics for the features provided in the Matlab .mat file:
`HW02knnFeatureSelectionBlindTestFeatures.mat`

Save the decision statistics to a Matlab `.mat` file, with the decision statistics stored in the vector `decStat` and saved in the same order as the blind test features. Submit the Matlab `.mat` file containing your decision statistics as an Attachment to the Assignment in Sakai.

(We know the corresponding targets for the blind test data, and will score your decision statistics to generate an ROC curve to evaluate your decision statistics.)