Lab02 - Questions and exercises

Notebook:	Didattica
Created:	October 27, 2014 11:11:30 AM
Updated:	November 3, 2014 5:51:42 PM
Author:	aittalam

(1) This is a variation of question 2, Section 2.4, page 52 on the course book.

Explain whether each scenario is **classification/regression/cluster analysis,** and indicate whether we are most interested in **prediction or inference**. Finally, provide *n* and *p.*

- we collect movie data from Netflix. From a set of 2000 movies we got genre, year, director, budget, user rating (1 to 5 stars), and we would like to use this information to predict how many stars a new movie will get

- we are analyzing an email mailbox with 10000 messages and would like to understand whether a new incoming message will be spam or not. For each message we have the sender, the subject, the date, and the body of the email, plus a class assignment (SPAM/NOT SPAM)

- we would like to know more about the types of users who visit a given website. From a set of 200000 users, we get user age, gender, and country, plus the time she spent on each of the 40 pages of the website.

(2) This is a variation of question 4, Section 2.4, page 53 on the course book.

Think about some real-life applications of statistical learning (pick up the exercise at page 53 and discuss together, plus comment the answers provided at https://github.com/asadoughi/stat-learning/blob/master/ch2/answers).

(3) Discuss the concepts of flexibility, bias/variance tradeoff, and draw together the 5 curves (as in Exercise 3a) in the flexibility/MSE plane. Start from the following main definitions and formulas.

(general) relationship between X, our *predictors*, and Y, our *responses*:

$$Y = f(X) + \epsilon.$$

Expected error:

$$\begin{aligned} E(Y-\hat{Y})^2 &= E[f(X)+\epsilon-\hat{f}(X)]^2 \\ &= \underbrace{[f(X)-\hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{Irreducible}} \end{aligned}$$

Mean Squared Error:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2,$$

Decomposition of the expected test MSE in Variance, Bias, and irreducible error:

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon)$$

---

(4) KNN exercise(s).

First, solve Exercise 7 on the book (page 53). Then, try the following ones on R.

```
% install library "animation"
install.packages("animation")

% generate a "training" dataset (12 points, two coordinates, tentatively two concentric circles drawn by hand)
train = matrix (c(2,2,2.5,2.7,3,1,1,2,2,3,4,4,2,3,3.5,2,3,2,3,4,1,1,1.5,4),12,2)

% test, using knn.ani, whether point (2.5,2.5) belongs to class "green" or "red"
knn.ani(train, c(2.5,2.5), c(rep("Red",5),rep("Green",7)), k = 3)

% verify by checking the distances
a = rowSums((x - train)^2)
min(a)
which.min(a)
a == min(a)

%% a more complex example, with randomly generated data
% generate 400 random points and divide them into two columns (to be used as random x,y coords)
train = matrix(rnorm(400,10,10),200,2)

% generate an index identifying points within a circle
idx = sqrt((train[,1]-10)^2+(train[,2]-10)^2)<5

% divide points into two classes (Red and Green)
classes = rep('Red',200)
classes[idx]='Green'

% generate 20 random points to be tested with KNN
test = matrix(rnorm(20,10,5),10,2)

% run the KNN animation using k = 3. TRY OTHER VALUES FOR K AND SEE WHAT HAPPENS!
knn.ani(train, test, classes, k = 3)
```

---

(5) Catch up with the last exercise of the previous class (on the "Auto" dataset), recap and show new commands

```
Auto = read.table('Auto.data',header=T,na.strings='?')
Auto = na.omit(Auto)
dim(Auto)
attach(Auto)
summary(Auto)
range(mpg)
```