582631 Introduction to Machine Learning, Fall 2016
Exercise set 2

Pen-and-paper problems

## Problem 1

*When the number of features p is large, there tends to be a deterioration in the performance of KNN and other local approaches that perform prediction using only observations that are near the test observation for which a prediction must be made. This phenomenon is known as the curse of dimensionality, and it ties into the fact that curse of dimensionality non-parametric approaches often perform poorly when p is large. We will now investigate this curse.*

(a) *Suppose that we have a set of observations, each with measurements on p = 1 feature, X. We assume that X is uniformly (evenly) distributed on [0, 1]. Associated with each observation is a response value. Suppose that we wish to predict a test observation's response using only bservations that are within 10% of the range of X closest to that test observation. For instance, in order to predict the response for a test observation with X = 0.6, we will use observations in the range [0.55, 0.65]. On average, what fraction of the available observations will we use to make the prediction?*

**ANSWER:**

If $x \in [0.05, 0.95]$ then the observations we will use are in the interval $[x-0.05, x+0.05]$ and consequently represents a length of 0.1 which represents a fraction of 10% . If $x<0.05$, then we will use observations in the interval $[0, x+0.05]$ which represents a fraction of $(100x+5)\%$ .

By a similar argument we conclude that if $x>0.95$, then the fraction of observations we will use is $(105-100x)\%$ . To compute the average fraction we will use to make the prediction we have to evaluate the following expression

$$\int_{0.05}^{0.95} 10dx + \int_{0}^{0.05} (100x+5)dx + \int_{0.95}^{1} (105x-100)dx = 9 + 0.375 + 0.375 = 9.75$$

So we may conclude that, on average, the fraction of available observations we will use to make the prediction is 9.75% .

(b) *Now suppose that we have a set of observations, each with measurements on p = 2 features, X1 and X2. We assume that (X1,X2) are uniformly distributed on [0, 1] × [0, 1]. We wish to predict a test observation's response using only observations that are within 10% of the range of X1 and within 10% of the range of X2 closest to that test observation. For instance, in order to predict the response for a test observation with X1 = 0.6 and X2 = 0.35, we will use observations in the range [0.55, 0.65] for X1 and in the range [0.3, 0.4] for X2. On average, what fraction of the available observations will we use to make the prediction?*

**ANSWER:**

If we assume $X\,1$ and $X\,2$ to be independant, the fraction of available observations we will use to make the prediction is $9.75\% \times 9.75\% = 0.950625\%$ .

(c) *Now suppose that we have a set of observations on p = 100 features. Again the observations are uniformly distributed on each feature, and again each feature ranges in value from 0 to 1. We wish to predict a test observation's response using observations within the 10% of each feature's range that is closest to that test observation. What fraction of the available observations will we use to make the prediction?*

**ANSWER:**

With the same argument than (a) and (b), we may conclude that the fraction of available observations we will use to make the prediction is $9.75\%^{100} \simeq 0\%$.

*(d) Using your answers to parts (a)–(c), argue that a drawback of KNN when p is large is that there are very few training observations "near" any given test observation.*

**ANSWER:**
As we saw in (a)-(c), the fraction of available observations we will use to make the prediction is $(9.75\%)^p$, where p is the number of features. So when $p \to \infty$ we have

$\lim_{p\to\infty}(9.75\%)^p = 0$.

*(e) Now suppose that we wish to make a prediction for a test observation by creating a p-dimensional hypercube centered around*
*the test observation that contains, on average, 10% of the training observations. For p = 1, 2, and 100, what is the length of each side of the hypercube? Comment on your answer.*

*Note: A hypercube is a generalization of a cube to an arbitrary number of dimensions. When p = 1, a hypercube is simply a line segment, when p = 2 it is a square, and when p = 100 it is a 100-dimensional cube.*

**ANSWER:**

For $p=1$ we have $l=0.1$ , for $p=2$ , we have $l=0.1^{1/2}$ and for $p=100$ , we have $l=0.1^{1/100}$ .

## Problem 2

## Exercise 7 (\Suppose that we wish to predict ...") on p. 170 of the course book.

*Suppose that we wish to predict whether a given stock will issue a dividend this year ("Yes" or "No") based on X, last year's percent profit.We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was $\bar{X} = 10$, while the mean for those that didn't was $\bar{X} = 0$. In addition, the variance of X for these two sets of companies was $\hat{\sigma}^2 = 36$. Finally, 80% of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue*
*a dividend this year given that its percentage profit was X = 4 last year.*

*Hint: Recall that the density function for a normal random variable*
*is $f(x) = \sqrt{1}$*
*$2\pi\sigma^2\, e^{-(x-\mu)^2/2\sigma^2}$*
*. You will need to use Bayes' theorem.*

**ANSWER:**

We have to plug in the parameters and X values in the equation for $p_k(x)$ (Eq. 4.12 from the book, page 139). Here we have:

X=4
$\pi_k$ = 80% =0.8
$\pi_{l=1-k}$ = 20% =0.2

$$\frac{1}{2\sigma^2} = 1/72$$

$\mu_k$=10

$\mu_{l=1-k}$=0

The term $\frac{1}{\sqrt{2\pi}\sigma}$ disappear because it appears in both numerator and denominator.

We get

$$p_1(4) = \frac{0.8e^{-(1/72)(4-10)^2}}{0.8e^{-(1/72)(4-10)^2} + 0.2e^{-(1/72)(4-0)}} = 0.752$$

so the probability that a company will issue a dividend this year given that its percentage return was X=4 last year is 0.752.

# Computer problems

## Problem 3 (3+3+3 points)

Even though the library class in R provides a ready-made implementation of the k-NN classifier, you get to do it yourself in this exercise (Yay!).

(a) (3 points) Download the classic MNIST handwritten digit database from http://yann.lecun.com/exdb/ mnist/, and load the data into R. [1] Display the fth training data instance on the screen to make sure you have succeeded. It should look more or less like a '9' (or a letter `a' leaning to the right but these are all supposed to be digits 0{9). Verify the that the correct class value, y, of the fth training instance is indeed 9 by printing the value train$y[5].

(b) (3 points) Use the rst 5 000 training instances and the rst 1 000 test instances only, and discard the rest.
(Unless you have a supercomputer or very much patience.) Compute all pairwise Euclidean distances,
$d(x_i; x_j) = \sqrt{\sum_{k=1}^{784}(x_{ik} - x_{jk})^2}$, where i runs through the 5 000 training instances, and j runs through the 1 000 test instances. Verify that the distance between the rst training instance and the rst test instance equals about 2395:8. Hint: Function dist in library proxy [2] does this very nicely but you can also write for loops.

(c) (3 points) Having stored the pairwise distances into a 5 000 1 000 distance matrix so that you don't have to recalculate them again later, classify each test instance by nding the k training instances nearest to it, and choosing the majority class among them. [3] Compute and plot the test set accuracy of the k-NN classi er with k = 1; : : : ; 50.

(continued on the next page...)

[1] Brendan O'Connor has kindly written a handy R script for reading the les: https://gist.github.com/brendano/39760. Just remember to put the les in folder mnist and unzip them. NB: Some systems may put a dot '.' in the le names where there should be a dash '-'. If the data loading script complains, check that the le names in the script match the actual le names.

[2] You can install libraries using install.packages("proxy"), etc.

[3] Hint: Here's a way to get the most common entry in a list: names(sort(table(...), decreasing=TRUE))[1], where you should write the name of the list at ....

Problem 4 (3+3+3 points)

Exercise 10 (item a{h) (\This question should be answered using the Weekly data set, ...") on p. 171 of the book. Note that the Lab starting on p. 154 is helpful here.

(a) (3 points) items a{b of the exercise in the book.

(b) (3 points) items c{d of the exercise in the book.

(c) (3 points) items e{h of the exercise in the book.