582631 Introduction to Machine Learning, Fall 2016
Exercise set 2

Due November 17th–18th. NB: Deadline for returning solutions my email (in case you can't attend a session) is 12:15 on Friday. Please attend the Thursday session(s) if possible. Friday tends to get crowded.

Continue reading the course book, pages 33–42 and 127–142.

Pen-and-paper problems

Problem 1 (3 points)

Exercise 4 ("When the number of features p is large, ...") on p. 168 of the course book.

Problem 2 (3 points)

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

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 fifth 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 fifth training instance is indeed 9 by printing the value `train$y[5]`.

(b) (3 points) Use the first 5 000 training instances and the first 1 000 test instances only, and discard the rest. (Unless you have a supercomputer or very much patience.) Compute all pairwise Euclidean distances, $d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^{784}(x_{ik} - x_{jk})^2}$, where i runs through the 5 000 training instances, and j runs through the 1 000 test instances. Verify that the distance between the first training instance and the first 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 finding 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 classifier with $k = 1, \ldots, 50$.

(continued on the next page...)

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