

# Machine Learning: Homework Assignment 3

E4525 Spring 2018,  
IEOR, Columbia University

Due: February 9th, 2018

1. **Choice of Error Function** In this problem we investigate how changes on error measure affect learning.

Assume we have  $N$  data points  $y_1 \leq y_2 \leq \dots \leq y_N$  and we want to find a “representative” value  $\hat{h} \in \mathbb{R}$ .

- (a) Show that if the loss function is the mean square error

$$E_2(h, \{y_i\}) = \frac{1}{N} \sum_i^N (h - y_i)^2 \quad (1)$$

the hypothesis that minimizes  $E_2$  is the sample mean

$$\hat{h}_2 = \frac{1}{N} \sum_i^N y_i \quad (2)$$

- (b) Show that if the loss function is the mean absolute error

$$E_1(h, \{y_i\}) = \frac{1}{N} \sum_i^N |h - y_i| \quad (3)$$

the hypothesis minimizing error is the sample median  $\hat{h}_1$  defined as

$$\sum_i \mathbb{1}(y_i \leq \hat{h}_1) = \frac{N}{2} \quad (4)$$

(do not worry about ties when  $N$  is even, etc)

- (c) Suppose  $y_N$  is perturbed to  $y_N + \epsilon$ , where  $\epsilon \rightarrow \infty$ . That single point has become an *outlier*. What happens our two estimators: the mean  $\hat{h}_2$  and the median  $\hat{h}_1$ ?

A health insurance company need to choose between two diagnosis procedures  $D_1$  and  $D_2$ .

Loss Function for Binary Classification

- Diagnosis  $D_1$  has a false negative rate (fail to diagnose a sick patient) of 20%, and a false positive rate (diagnose as sick a healthy patient) of 1%.

- Diagnosis  $D_2$  has a false positive rate of 10% and a false negative rate of 5%

The cost of the company of a false negative (diseases goes untreated) is \$1,000, and a false positive (unnecessary further testing of a healthy subject) is \$10.

The prevalence of the disease of the population is 1%.

- What is the error rate (number of incorrectly classified patients) of each method?
  - What are the expected costs of each method? Which test will the company choose?
  - For the patient, an extra visit to the doctor in the case of a false positive is a big hassle. Lets say his cost is \$100 for a false positive and still \$1,000 for a false negative. What are the expected costs to patients of  $D_1$  and  $D_2$ ? Which one would they prefer?
2. Considering only linear combinations of monomials  $x^k$  for  $k = 1, \dots, K$ , describe a good hypothesis space to approximate a continuous function  $f(x)$  defined in  $(-1, 1)$  given that we have the following constraints:
- No constraints:  $f(x)$  is arbitrary.
  - $f$  is even:  $f(-x) = f(x)$ .
  - $f$  is odd:  $f(-x) = -f(x)$ .
  - $f(-1) = 0$  and  $f(1) = 0$ .
  - $f(0) = 1$ .
- 4
- Assume a function  $f(x)$  is even. Which one of the hypothesis spaces will produce a lower training error 2a or 2b? Which one do you expect to have lower out of sample test error? Explain your reasoning.