Machine Learning: Homework Assignment 3

E4525 Spring 2018,
IEOR, Columbia University

Due: February 15th, 2019

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

   Assume we have N data points $y_1 \leq y_2 \leq \cdots \leq y_N$ 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

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

   the hypothesis that minimizes E_2 is the sample mean

   $$\hat{h}_2 = \frac{1}{N} \sum_i^N y_i \tag{2}$$

   Setting the gradient of E_2 to zero

   $$0 = \frac{\partial E_2}{\partial h} = \frac{2}{N} \sum_i^N (h - y_i) \tag{3}$$

   Solving trivially gets equation 2

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

   $$E_1(h, \{y_i\}) = \frac{1}{N} \sum_i^N |h - y_i| \tag{4}$$

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

   $$\sum_i \mathbb{1}(y_i \leq \hat{h}_1) = \frac{N}{2} \tag{5}$$

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

Setting to zero the gradient again

$$0 = \frac{\partial E_1}{\partial h} = \frac{2}{N} \sum_i^N \text{sgn}(h - y_i) \tag{6}$$

where $\text{sgn}(z)$ is the sign function

$$\text{sgn}\, z = \begin{cases} -1 & \text{for } z < 0 \\ 1 & \text{for } z > 0 \end{cases} \tag{7}$$

The minimum of that expression will be when there are many y's to the left of h (contributing -1) as there are to the right (contribution +1). That is the definition of the median.

(c) Suppose y_N is perturbed to $y_N + \epsilon$, where $\epsilon \to \infty$. That single point has become an *outlier*. What happens our two estimators: the mean $\hat{h}_2$ and the median $\hat{h}_1$?
$\hat{h}_2 \to \infty$.
$\hat{h}_1$ changes by at most one position on the list $\{y_i\}$ (the median is a robust estimator).

2. **Loss Function for Binary Classification** A health insurance company need to choose between two diagnosis procedures D_1 and D_2.

- 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 negative rate of 10% and a false positive 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%.

(a) What is the error rate (number of incorrectly classified patients) of each method?

$$\text{Error} = \text{FN} * P_{\text{sick}} + \text{FP} * (1 - P_{\text{sick}}) \tag{8}$$

- D_1: 20% * 1% + 1% * 99% = 1.19%
- D_2: 10% *1% + 5% * 99% = 5.05%

(b) What are the expected costs of each method? Which test will the company choose?

$$\text{Cost} = C_{\text{FN}} * \text{FN} * P_{\text{sick}} + C_{\text{FP}} * \text{FP} * (1 - P_{\text{sick}}) \tag{9}$$

- D_1: 1,000*20% * 1% + 10*1% * 99% = 2.1
- D_2: 1,000*10% *1% + 10*5% * 99% = 1.5

Company would prefer D_2 even when it is less accurate.

(c) 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?

$$\text{Cost}_{\text{Patient}} = C^{\text{P}}_{\text{FN}} * \text{FN} * P_{\text{sick}} + C^{\text{P}}_{\text{FP}} * \text{FP} * (1 - P_{\text{sick}}) \tag{10}$$

- D_1: 1,000*20% * 1% + 100*1% * 99% = 2.99
- D_2: 1,000*10% *1% + 100*5% * 99% = 5.95

Patients would prefer D_1 even though it is more costly for the health company.

3. **Hypothesis Spaces** Considering only linear combinations of monomials x^k for $k = 0, \dots K$, describe a good hypothesis space to approximate a continuous function $f(x)$ defined in $(-1, 1)$ given that we have the following constrains:

(a) No constrains: $f(x)$ is arbitrary.
Arbitrary polynomials

$$h(x) = \sum_{k=0}^{K} w_k x^k \tag{11}$$

(b) f is even: $f(-x) = f(x)$.
Even polynomials

$$h(x) = \sum_{k=0}^{K/2} w_k x^{2k} \tag{12}$$

(c) f is odd: $f(-x) = -f(x)$.
Odd polynomials

$$h(x) = \sum_{k=0}^{K/2} w_{2k+1} x^{2k+1} \tag{13}$$

(d) $f(-1) = 0$ and $f(1) = 0$.
Polynomials with a $(x-1)(x+1)$ factor

$$h(x) = (x-1)(x+1) \sum_{k=0}^{K-2} w_k x^k \tag{14}$$

(e) $f(0) = 1$.

$$h(x) = 1 + x \sum_{k=0}^{K-1} w_k x^k \tag{15}$$

(f) Assume a function $f(x)$ is even. Which one of the hypothesis spaces will produce a lower training error 3a or 3b? Which one do you expect to have lower out of sample test error? Explain your reasoning.

- **Training error** 3a will have a lower training error.
  It includes all the functions in 3b plus all the odd polynomials.
- **Test Error** 3b will have lower test error.
  The bias is the same on both spaces because only even powers contribute to the true function. The variance will be larger on 3a because it has extra functions (odd powers) that do not contribute to reduce bias, they will just **overfit** the data.