# Machine Learning
# Assignment 5

Mao Xiqing(tls868)

January 5, 2021

## Contents

# 1 A bit more on the VC-dimension

1. From the definition we know $d_{VC}$ is the largest sample size which can be shattered by $\mathcal{H}_d$:
$$d_{VC}(\mathcal{H}_d) = max\{n|m_{\mathcal{H}_d}(n) = 2^n\}. \tag{1}$$
For a binary decision trees with depth $d$, $2^d$ samples can be shattered. So $d_{VC}(\mathcal{H}_d) = 2^d$.

2. For a binary decision trees with infinite depth, one can feed arbitrarily many samples to the tree, and still classify all of them, which means $m_{\mathcal{H}_{d(n)}} = 2^n$ for all $n$, then $d_{VC}(\mathcal{H}_d) = \infty$.

# 2 Separating Hyperplanes

We know: sample size $n = 100000$, margin $1/\|w\| = 0.1$, confidence $\delta = 0.01$ and $\hat{L}_{FAT}(h, S) = 0.01$. Put them into Theorem 3.22:

$$\mathbb{P}\left(\exists h \in \mathcal{H} : L_{\mathrm{FAT}}(h) \leq \hat{L}_{\mathrm{FAT}}(h, S) + \sqrt{\frac{8\ln\left(2\left((2n)^{1+\lceil\|\mathrm{w}\|^2\rceil} + 1\right)\left(1 + \lceil\|\mathrm{w}\|^2\rceil\right)\lceil\|\mathrm{w}\|^2\rceil/\delta\right)}{n}}\right) \geq 1 - \delta \tag{2}$$

we can get a bound on its expected loss that holds with probability of 99%:

$$0.01 \leq L_{FAT}(h) \leq 0.3259 \tag{3}$$

# 3 The fine details of the lower bound

We want there exist a distribution $p(X, Y)$ and a hypothesis space $\mathcal{H}$ with infinite VC dimension, such that for any sample S of more than 100 points are satisfy

$$\mathbb{P}\left(L(h) \leq \hat{L}(h, S) + 0.01\right) \geq 0.95. \tag{4}$$

Let $p(X, Y) = \{(x_1, y_1), (x_2, y_2) \ldots, (y_n, y_n)\}$, $p(x_i = a) = 1, q(y_i = 0) = 1$. Then the empirical loss $\hat{L}(h, S)$:

$$\hat{L}(h, S) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i), y_i) = \frac{1}{100} \times 100\ell(h(1), 0) = \ell(h(1), 0), \tag{5}$$

and the expected loss $L(h)$:

$$L(h) = \mathbb{E}[\ell(h(x_i), y_i)] = \mathbb{E}[\ell(h(1), 0)] = \ell(h(1), 0), \tag{6}$$

which means $L(h) = \hat{L}(h, S) = \ell(h(1), 0)$, now the inequality holds.

# 4 Random Forests

## 4.1 Normalization

Is nearest neighbor classification affected by this type of normalization? **Yes.** Is random forest classification affected by this normalization? **No.**

In nearest neighbor classification, the decisions are made by Euclidean distance:

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}. \tag{7}$$

If $\text{feature}_1 \in [0, 9999]$ while $\text{feature}_2 \in [0, 50]$, it is obvious that $\text{feature}_1$ will contribute more distance. By normalization, different features have the same scale. In this way, when using algorithm to learn parameters, different features have the same influence on the parameters.

However, in decision tree (random forest), the scaling of input data is not a determinant of the predictions, because it based on square error or impurity measure (Gini index) rather than distance, which related to mean and probability, respectively. And scaling does not affect the order of features, the location of the split point nor the structure of the tree. So a transformation of the input will not affect random forest.

## 4.2 Random forests in practice

Source code for this question can be found in the notebook `rf.ipynb`.

I tried to use GridSearchCV to find some parameters like `max_depth`, but it seems make no sense.

Test error: 0.1031.

I use following code to define the classifier:

```
model = RandomForestClassifier(n_estimators = 50, n_jobs = -1, oob_score
    = True, random_state=60)
```