Topics: PAC learning
- Support Vector Machines (SVMs)

Overview
- ERH principle: minimize the errors on the training set
- Linear predictors: \( \hat{y} = \text{sign}(w \cdot x) \)
- Perceptron: linear predictor + ERH principle
- Bayes optimal predictor: \( P(y = 1|x) > \frac{1}{2} \Rightarrow \hat{y} = +1 \)
- Maximum likelihood estimator of \( P(x|y) \)

PAC Learning

1) How can we define what type of a hypothesis class is suitable for learning?
2) It might be the case that the training set of example is not representative.

Denote by \( \delta \) the probability of getting a non-representative sample \( \Rightarrow \) we cannot guarantee perfect label prediction

Denote by \( \delta \) the accuracy of the predictor:
\[
\text{if } \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(y_i \neq f(x_i)) > \delta \Rightarrow \text{failure of the learner}
\]
\[
\text{if } \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(y_i = f(x_i)) < \delta \Rightarrow \text{we view the output of the algorithm as an approximately correct predictor}
\]
Definition

A hypothesis class $H$ is PAC learnable with respect to a loss function $L(x, y; \theta)$, if there exist a minimal number of training examples $m$ drawn iid from distribution $D$ such that

$$\left| E_x \left[ \ell(x, y; \theta) \right] \right| \leq \min_{\theta'} \left| E_{x,y} \left[ \ell(x, y; \theta') \right] \right| + \varepsilon$$

with probability $1 - \delta$ over the choice of the $m$ training examples.

Finite hypothesis classes are PAC learnable.

If the set of hypotheses is finite (finite number of possible functions) then we can show that they can be PAC learnable:

Chebyshev's inequality

For random variable $Z$, $\varepsilon > 0$:

$$P\left\{ |Z - E[Z]| > \varepsilon \right\} \leq \frac{\text{Var}(Z)}{\varepsilon^2}$$

There is better inequality; that works on a sequence of RV:

Hoeffding’s inequality

Let $Z_1, Z_2, \ldots, Z_n$ be a sequence of iid random variables, where $a < Z_i < b$ for all $i$, and $E[Z_i] = \mu$. Then for $\varepsilon > 0$:

$$P\left\{ \frac{1}{n} \sum_{i=1}^{n} Z_i - \mu > \varepsilon \right\} \leq 2 \exp \left( -\frac{2\varepsilon^2}{(b-a)^2} \right)$$
Proof that finite clones are PAC learnable:

We need to find $m$ such that

$$IP_s \left( \forall h \in H, \left| \frac{1}{m} \sum_{i=1}^{m} \ell(x_i, y_i; h) - \mathbb{E}[\ell(x, y; h)] \right| \leq \varepsilon \right) \geq 1 - \delta$$

or

$$IP_s \left( \exists h \in H, \left| \frac{1}{m} \sum_{i=1}^{m} \ell(x_i, y_i; h) - \mathbb{E}[\ell(x, y; h)] \right| > \varepsilon \right) < \delta$$

$$IP_s \left( \exists h \left| \frac{1}{m} \sum_{i=1}^{m} \ell - \mathbb{E}[\ell] \right| > \varepsilon \right) < \sum_{h \in H} IP_s \left( \left| \frac{1}{m} \sum_{i=1}^{m} \ell - \mathbb{E}[\ell] \right| > \varepsilon \right)$$

Hoeffding

$$\ell \in [0, 1] \quad \Rightarrow \quad \sum_{h \in H} \mathbb{P} \left( \left| \frac{1}{m} \sum_{i=1}^{m} \ell - \mathbb{E}[\ell] \right| > \varepsilon \right) \leq 2 \exp \left\{ -2m \varepsilon^2 \right\}$$

$$= 2 \log (2H) \exp \left\{ -2m \varepsilon^2 \right\} = \delta$$

$$\Rightarrow \quad m \geq \frac{\log (2H) / \delta}{2\varepsilon^2}$$
Support Vector Machine (SVM)

Perceptron algorithm can cause overfitting. Assume linear separable data. The following are solutions to Perceptron:

These solutions which are typical to Perceptron are overfitting since they are tight to the data points: if we want to infer the examples marked as ■ we will probably fail!

Alternatively, we want a solution just in the middle between classes.
Specifically, we would like to define it as follows:

1) Classify the data correctly \( wX_i > 1 \) for \( y_i = +1 \)
   and \( wX_i < -1 \) for \( y_i = -1 \)
   or equivalently:
   \[
y_i wX_i \geq +1
   \]

2) Maximum margin. The margin size, when \( x_1, x_2 \) are two points on the margin, is:

\[
\frac{w}{||w||} (x_2 - x_1)
\]

Since \( |wX_1| = 1 \) and \( |wX_2| = 1 \), we have the size of the margin is

\[
\frac{1 - (-1)}{||w||} = \frac{2}{||w||}
\]

Combining (1) and (2) together:

\[
W^* = \arg \max_w \frac{2}{||w||} \text{ such that } y_i wX_i \geq 1 \text{ for all } i
\]

which is the same as

\[
W^* = \arg \min_w \frac{1}{2} ||w||^2 \text{ such that } y_i wX_i \geq 1 \text{ for all } i
\]