1 Section 1 (Apr. 5th)

1.1 Binary classification on R

- Domain: \( R \).
- Target concept: \( h_t(x) \) \((t \in R)\)
  \[
  h_t(x) = \begin{cases} 
  +1 & \text{if } x \leq t \\
  -1 & \text{if } x > t 
  \end{cases}
  \]

- Hypothesis space: \( H = \{ h_k(x) | k \in R \} \)
- Version space: Let \( S \) be \( N \) samples \( \{x_1, x_2, \ldots, x_N\} \) from \( R \) paired with their labels \( h_t(x_i) \):
  \[
  S = \{ [x_1, h_t(x_1)], [x_2, h_t(x_2)], \ldots [x_n, h_t(x_n)] \} \quad (x_i \in R)
  \]
  Let \( x_{\max}^+ \) and \( x_{\min}^- \) be the largest sample with ”+” label and smallest sample with ”-” label in \( S \). Then the version space of \( S \) is:
  \[
  V(S) = \{ h_k(x) | x_{\max}^+ \leq k \leq x_{\min}^- \}
  \]
  And the most specific one is \( h_{x_{\max}^+} \).

1.2 Generalization error

1.2.1 Generalization error as a random variable

Given a domain \( X \) and a target concept \( h_t \), the generalization error of a hypothesis \( h_t \) is the probability of the event \( h_t(x) \neq h_l(x) \) when \( x \) is drawn from \( X \) according to some distribution \( D \):

\[
Pr_D(h_t(x) \neq h_l(x))
\]

For a given number \( \epsilon \), we say a hypothesis \( h_t \) is \( \epsilon - \text{good} \) if the generalization error of \( h_t \) is equal to or smaller than \( \epsilon \).
The generalization error of a learning algorithm is the error of its learned hypothesis after seeing some training data. In most cases, the training data is sampled independently from the domain using the same distribution $D$ as we measure the error (it will be referred as sampling distribution $D$ regardless of training or testing.) So the generalization error, which is a function of the training data, is also a random variable on interval $[0,1]$. In general, there are three quantities in our concern, the number of samples $N$, the error $\epsilon$ and the probability of the learning algorithm will return a $\epsilon$-good hypothesis.

1.2.2 Probably Approximately Correcting (PAC) Learning

PAC learning theory is a method to model the relation of these three quantities. For a learning algorithm, we want to know: given any sampling distribution $D$ and some constant pair $(\epsilon, \delta)$, how many training samples are needed to guarantee that the algorithm will learn a $\epsilon$-good hypothesis with probability at least $1 - \delta$.

For the classification problem discussed above, the learning algorithm always return a hypothesis $h_l$ with $l \leq t$. The generalization error of $h_l$ is the probability of a sample drawn from interval $[l, t]$. Let $f$ be the density function of sampling distribution $D$. This probability $f([l, t])$ is a monotone decrease function when $l$ moves towards the target concept $t$. So for a desired error $\epsilon_0$, $\epsilon_0$-good hypotheses lies between $t$ and $l_0$ where $f([l_0, t]) = \epsilon_0$. Since the algorithm always return the most specific hypothesis, the probability of returning a good hypothesis is the probability of seeing a training sample in the interval $[l_0, t]$. Given training samples are independently drawn with $D$, this probability is $1 - (1 - f(l_0, t))^N = 1 - (1 - \epsilon_0)^N$.

Using the bound $1 - x \leq e^{-x}$, we may restate this as for given error $\epsilon_0$, $N$ training samples have at least $1 - e^{-\epsilon_0 N}$ probability to return a good hypothesis. In other words, to achieve error $\epsilon_0$ with probability $1 - \delta$, we need at most $\frac{1}{\epsilon_0} \ln(\frac{1}{\delta})$ training samples.

You may refer to the following lecture note for a detailed discussion:
(http://www.cs.princeton.edu/courses/archive/spr06/cos511/lecture_notes/0214.pdf)

probabilities of the label + to

2 Section 2 (Apr. 12th)

2.1 Calculate conditional probabilities in a group of random variables

- Identify random variables

Example: Let the domain (instance space) have 2 points: $a$ and $b$. Consider the model where first an hypothesis is chosen from $\{h_1, h_2\}$ and then instances ($a$ and $b$’s) are labeled according to the chosen hypothesis. One sample is labeled: $(a; +)$.

There are 2 random variables in this example: $h \in \{h_1, h_2\}$, $a_1 \in \{+, -\}$. 

2
Example: Three samples are labeled: (a; +), (a; -) and (b; +).

There are 4 random variables in these example: \( h \in \{ h_1, h_2 \} \), \( a_1 \in \{ +, - \} \), \( b_1 \in \{ +, - \} \) and \( a_2 \in \{ +, - \} \)

- Identify the joint distribution by generating these random variables

Example: The prior distribution of \( h \) are \( P(h_1) = \frac{1}{3} \), \( P(h_2) = \frac{2}{3} \). \( h_1 \) and \( h_2 \) assign the label + to each instance with the following probabilities (the probability of the label − is 1 minus the probability of +).

\[
\begin{array}{cc}
  h_1 & h_2 \\
    a & 0.8 & 0.25 \\
    b & 0.4 & 0.75
\end{array}
\]

Joint distribution in two random variables case:

\[
P(h, a_1) = P(h)P(a_1|h)
\]

\[
P(a_1 = +|h_1) = 0.8 \quad P(a_1 = -|h_1) = 1 - 0.8 = 0.2
\]

Joint distribution in four random variables case:

\[
P(h, a_1, b_1, a_2) = P(h)P(a_1|h)P(a_2|h)P(b_1|h)
\]

\[
P(a_2|h) = P(a_1|h)
\]

\[
P(b_1 = +|h_1) = 0.4 \quad P(b_1 = -|h_1) = 1 - 0.4 = 0.6
\]

- Calculate distributions

Example: What is the prior distribution of + for point a before seeing the data?

Prior distribution is the distribution without any condition.

\[
P(a_1) = P(a_1, h = h_1) + P(a_1, h = h_2)
\]

Example: What is the posterior distribution of hypothesis after seeing three labeled samples? What is the maximum a posteriori hypothesis?

“after seeing three labeled samples” means the distribution conditioned on the sample random variables with observed values.

\[
P(h = h_1|a_1 = '+', a_2 = '-', b_1 = '+')
\]

\[
= \frac{P(h = h_1, a_1 = '+', a_2 = '-', b_1 = '+')}{P(a_1 = '+', a_2 = '-', b_1 = '+')}
\]

\[
= \frac{P(h = h_1, a_1 = '+', a_2 = '-', b_1 = '+')}{\sum_{h \in \{h_1, h_2\}} P(h, a_1 = '+', a_2 = '-', b_1 = '+')}
\]
Example: What is the (mean) posterior (full Bayesian) probabilities of + for b after seeing the three samples?

Add a new random variable $b_2$ and the new joint distribution is:

$$P(h, a_1, b_1, a_2, b_2) = P(h)P(a_1|h)P(a_2|h)P(b_1|h)P(b_2|h)$$

$$P(b_2|h) = P(b_1|h)$$

The conditional distribution

$$P(b_2|a_1 = '+', b_1 = '-' , a_2 = '+')$$

$$= \frac{P(b_2, a_1 = '+', b_1 = '-', a_2 = '+')} {P(a_1 = '+', b_1 = '+', a_2 = '-')}$$

$$= \frac{\sum_{h \in \{h_1, h_2\}} P(b_2, h, a_1 = '+', b_1 = '+', a_2 = '-')} {\sum_{b_2 \in \{+,-\}} \sum_{h \in \{h_1, h_2\}} P(b_2, h, a_1 = '+', b_1 = '+', a_2 = '-')}$$