Tutorial 7 : Supervised Learning

1 The Perceptron Algorithm

A set of labeled samples $\{x_i, y_i\}_{i=1}^n$ is given, where $x_i \in \mathbb{R}^m$ and $y_i \in \{-1, 1\}$.

Goal: Find a linear classifier $f : \mathbb{R}^m \to \{-1, 1\}$ which satisfies

$$f(x_i) = sign(w^T x_i) = y_i,$$

where $w \in \mathbb{R}^m$ is a vector of weights.

The Algorithm

- Initialization - set initial weight vector w_0.
- For $t = 1, 2, ...$
  1. Pick a sample $\{x_t, y_t\}$ from the training set
  2. Compute
  $$\hat{y}_t = sign(w_t^T x)$$
  3. Update the weight vector
  $$w_{t+1} = w_t + \frac{1}{2}(y_t - \hat{y}_t)x_t$$

The algorithm converge in finite number of iterations if the problem is linear separable.
In the linear non-separable case, there is no guarantee for convergence.

Question 1

In this exercise we aim to prove the convergence of perceptron algorithm when the set of examples is linear separable. Under this assumption, there exists a weight vector w^* for which

$$y_i \langle w^*, x_i \rangle \geq 1, \ i = 1, 2, ..., n.$$

Consider the following version of the perceptron algorithm

- **Input:** $\{x_i, y_i\}_{i=1}^n$, $x_i \in \mathbb{R}^m$, $y_i \in \{-1, 1\}$.
- **Initialization:** $w_1 = (0, ..., 0)$.
- **For** $t = 1, 2, ...$

  If $\exists i$ such that $y_i \langle w_t, x_i \rangle \leq 0$
  $$w_{t+1} = w_t + y_i x_i.$$
  Else, return w_t and finish.

(a) Prove that $\langle w^*, w_{T+1} \rangle \geq T$. Hint: Use a telescoping series for all iterations up to T.

(b) Define $R = \max_i ||x_i||_2^2$. Prove that $||w_{t+1}||_2^2 \leq ||w_t||_2^2 + R^2$.

(c) Show that $||w_{T+1}||_2^2 \leq TR^2$.

(d) We want to show that the algorithm converges to w^*, i.e.,

$$\cos\theta_{T+1} = \frac{\langle w^*, w_{T+1}\rangle}{||w^*||_2||w_{T+1}||_2} \underset{T\to\infty}{\to} 1.$$

Explain the geometric meaning of this condition.

(e) Define $B = \min\{||w|| : \; y_i\langle w, x_i\rangle \geq 1 \; \forall i \in [1, n]\}$ and let w^* be the vector which achieves this minimum. Use the previous parts to obtain a lower bound on $\cos\theta_{T+1}$. What is a trivial upper bound on $\cos\theta_{T+1}$?

(f) Use the bounds you found to prove the convergence of the algorithm. Find an upper bound on the number of iteration required until convergence.

Solution

(a) First notice that

$$\langle w^*, w_{t+1}\rangle - \langle w^*, w_t\rangle = \langle w^*, w_{t+1} - w_t\rangle = \langle w^*, y_i x_i\rangle = y_i\langle w^*, x_i\rangle \geq 1.$$

Hence,

$$\langle w^*, w_{T+1}\rangle = \sum_{t=1}^{T}\Big(\langle w^*, w_{t+1}\rangle - \langle w^*, w_t\rangle\Big) \geq \sum_{t=1}^{T} 1 = T.$$

(b) It holds that

$$\begin{aligned} y_i\langle w_t, x_i\rangle > 0 \; &\Rightarrow ||w_{t+1}||_2^2 = ||w_t||_2^2, \\ y_i\langle w_t, x_i\rangle \leq 0 \; &\Rightarrow ||w_{t+1}||_2^2 = ||w_t + y_i x_i||_2^2 = ||w_t||_2^2 + \underbrace{2\langle w_t, y_i x_i\rangle}_{\leq 0} + ||x_i||_2^2 \leq ||w_t||_2^2 + R^2. \end{aligned}$$

(c) Using that $w_0 = (0, ..., 0)$ and the previous part we have

$$||w_{T+1}||_2^2 \leq \sum_{t=1}^{T} R^2 = TR^2.$$

(d) In the limit $T \to \infty$, the vectors w_{T+1} and w^* have the same direction, therefore, they will classify the examples in the same manner with respect to their sign.

(e) A trivial upper bound for $\cos\theta_{T+1}$ is $\cos\theta_{T+1} \leq 1$. A lower bounds can be achieved using the previous parts

$$\cos\theta_{T+1} = \frac{\langle w^*, w_{T+1}\rangle}{||w^*||_2||w_{T+1}||_2} \geq \frac{T}{B\sqrt{TR^2}} = \frac{\sqrt{T}}{BR}.$$

Hence,

$$\frac{\sqrt{T}}{BR} \leq \cos\theta_{T+1} \leq 1.$$

(f) Notice that for $T = B^2R^2$ we get that

$$\cos\theta_{T+1} \geq \frac{\sqrt{T}}{BR} = \frac{\sqrt{B^2R^2}}{BR} = 1 \; \to \theta_{T+1} = 0,$$

which implies that w_{T+1} and w^* are aligned and the algorithm converged. Thus, an upper bound on the number of iteration is given by $(BR)^2$.

Naive Bayes Classifier

Notation

Ω - Output space : a finite set of classes $\omega_i \in \Omega,\ i = 1, 2, ..., N$.

X - Input space : $x \in X$.

f - A classifier $f : X \to \Omega$ which maps $x \in X$ to $\omega \in \Omega$.

Optimal Bayes Classifier

$$f(x) = \underset{i=1,2,\ldots,N}{\arg\max} \quad p(x|\omega_i)p(\omega_i).$$

Empirical Bayes classifier

A training set of labeled examples $\{x_k, y_k\}_{k=1}^{m}$ is given, where $x_k \in X$ and $y_k \in \Omega$.

1. Estimate the distributions $p(x|\omega)$ and $p(w)$ from the training set $\{x_k, y_k\}_{k=1}^{m}$.
2. Use the estimated distributions to compute the optimal Bayes classifier.

Naive Bayes classifier

When the dimension n of the input space is high, estimating $p(x|w)$ is complicated and in most cases not practical. One possible approach for dealing with this problem is to assume independence between the coordinates of the input $x = (x_1, x_2, ..., x_n)^T$, that is we make the naive assumption (hence the name) that

$$p(x|\omega) \approx \prod_{i=1}^{d} p(x_i|\omega).$$

Then, estimate the marginal one-dimensional distributions $\{p(x_i|\omega)\}_{i=1}^{d}$ using the training set.

Question 2

Consider the input vector to be $x = (x_1, x_2, ..., x_n)^T$ where $x_i \in \{0, 1\}$ and the output targets are a single binary-value $y \in \{0, 1\}$. Our model is then parameterized by

$$\begin{aligned}
p_1 &= p(y = 1), \\
q_i &= p(x_i = 1|y = 0),\ i = 1, 2, ..., n \\
h_i &= p(x_i = 1|y = 1),\ i = 1, 2, ..., n.
\end{aligned}$$

(a) Model the distibutions $p(y)$, $p(x|y = 0)$ and $p(x|y = 1)$ using $p_1, q_1, .., q_n.h_1, ..., h_n$.

(b) A labeled training set $\{x^{(k)}, y^{(k)}\}_{k=1}^{m}$ is given.
Find the joint likelihood function $\ell(\theta) = \log \prod_{k=1}^{m} p(x^{(k)}, y^{(k)}; \theta)$ where θ represents the entire set of parameters $\theta = \{p_1, q_1, .., q_n.h_1, ..., h_n\}$.

(c) Find the parameters which maximize the likelihood function.

(d) Consider making a prediction on some new data point x using the most likely class estimate generated by the naive Bayes algorithm. Show that the naive Bayes classifier is a linear classifier, i.e., if $p(y = 0|x)$ and $p(y = 1|x)$ are the class probabilities returned by naive Bayes, show that there exists some $u \in \mathbb{R}^{n+1}$ such that

$$p(y = 1|x) \geq p(y = 0|x) \ \Leftrightarrow\ u^T \begin{bmatrix} 1 \\ x \end{bmatrix} \geq 0.$$

Solution

(a) We model the distributions as follows

$$
\begin{aligned}
p(y) &= p_1^y (1-p_1)^{(1-y)}, \\
p(x|y=0) &= \prod_{i=1}^{n} p(x_i|y=0) = \prod_{i=1}^{n} q_i^{x_i}(1-q_i)^{1-x_i}, \\
p(x|y=1) &= \prod_{i=1}^{n} p(x_i|y=1) = \prod_{i=1}^{n} h_i^{x_i}(1-h_i)^{1-x_i}.
\end{aligned}
$$

(b) The joint likelihood function is given by

$$
\begin{aligned}
\ell(\theta) &= \log \prod_{k=1}^{m} p(x^{(k)}, y^{(k)}; \theta) \\
&= \log \prod_{k=1}^{m} p(x^{(k)}|y^{(k)}; \theta) p(y^{(k)}; \theta) \\
&= \log \prod_{k=1}^{m} \left(\prod_{i=1}^{n} p(x_i^{(k)}|y^{(k)}; \theta) \right) p(y^{(k)}; \theta) \\
&= \sum_{k=1}^{m} \left(\sum_{i=1}^{n} \log p(x_i^{(k)}|y^{(k)}; \theta) + \log p(y^{(k)}; \theta) \right) \\
&= \sum_{k=1}^{m} \Bigg(\log \left(p_1^{y^{(k)}} (1-p_1)^{(1-y^{(k)})} \right) + \sum_{i=1}^{n} y^{(k)} \log \left(h_i^{x_i^{(k)}} (1-h_i)^{1-x_i^{(k)}} \right) \\
&\qquad + \sum_{i=1}^{n} (1-y^{(k)}) \log \left(q_i^{x_i^{(k)}} (1-q_i)^{1-x_i^{(k)}} \right) \Bigg) \\
&= \sum_{k=1}^{m} \Bigg(y^{(k)} \log p_1 + (1-y^{(k)}) \log(1-p_1) + \sum_{i=1}^{n} y^{(k)} \Big(x_i^{(k)} \log h_i + (1-x_i^{(k)}) \log(1-h_i) \Big) \\
&\qquad + \sum_{i=1}^{n} (1-y^{(k)}) \Big(x_i^{(k)} \log q_i + (1-x_i^{(k)}) \log(1-q_i) \Big) \Bigg)
\end{aligned}
$$

(c) To find the parameters we set the gradient of $\ell(\theta)$ to zero -

$$
\begin{aligned}
\frac{\partial \ell}{\partial p_1} &= \sum_{k=1}^{m} y^{(k)} \frac{1}{p_1} - (1-y^{(k)}) \frac{1}{(1-p_1)} = 0 \\
\Leftrightarrow & \sum_{k=1}^{m} y^{(k)}(1-p_1) - (1-y^{(k)}) p_1 = 0 \\
\Leftrightarrow & \sum_{k=1}^{m} y^{(k)} = \sum_{k=1}^{m} p_1 \\
\Leftrightarrow & \; p_1 = \frac{1}{m} \sum_{k=1}^{m} y^{(k)} = \frac{1}{m} \sum_{k=1}^{m} 1\{y^{(k)} = 1\}
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial \ell}{\partial h_i} = \sum_{k=1}^{m} y^{(k)} \left(x_i^{(k)} \frac{1}{h_1} - (1 - x_i^{(k)}) \frac{1}{1 - h_1} \right) = 0 \\
&\Leftrightarrow \sum_{k=1}^{m} y^{(k)} \left(x_i^{(k)} (1 - h_1) - (1 - x_i^{(k)}) h_1 \right) = 0 \\
&\Leftrightarrow \sum_{k=1}^{m} y^{(k)} x_i^{(k)} = \sum_{k=1}^{m} y^{(k)} h_i \\
&\Leftrightarrow h_i = \frac{\sum_{k=1}^{m} y^{(k)} x_i^{(k)}}{\sum_{k=1}^{m} y^{(k)}} = \frac{\sum_{k=1}^{m} 1\{y^k = 1 \cap x_i^{(k)} = 1\}}{\sum_{k=1}^{m} 1\{y^k = 1\}}
\end{aligned}
$$

The solution for q_i proceeds in the identical manner:

$$
q_i = \frac{\sum_{k=1}^{m} (1 - y^{(k)}) x_i^{(k)}}{\sum_{k=1}^{m} (1 - y^{(k)})} = \frac{\sum_{k=1}^{m} 1\{y^k = 0 \cap x_i^{(k)} = 1\}}{\sum_{k=1}^{m} 1\{y^k = 0\}}
$$

(d) We will classify $y = 1$ if

$$
\begin{aligned}
&p(y = 1|x) \geq p(y = 0|x) \\
&\Leftrightarrow \frac{p(y = 1|x)}{p(y = 0|x)} \geq 1 \\
&\Leftrightarrow \frac{\prod_{i=1}^{n} p(x_i|y = 1) p(y = 1)}{\prod_{i=1}^{n} p(x_i|y = 0) p(y = 0)} \geq 1 \\
&\Leftrightarrow \log \frac{p(y = 1)}{p(y = 0)} + \sum_{i=1}^{n} \log \frac{p(x_i|y = 1)}{p(x_i|y = 0)} \geq 0 \\
&\Leftrightarrow \log \frac{p_1}{1 - p_1} + \sum_{i=1}^{n} x_i \log \frac{h_i}{q_i} + (1 - x_i) \log \frac{1 - h_i}{1 - q_i} \geq 0 \\
&\Leftrightarrow \log \frac{p_1}{1 - p_1} + \sum_{i=1}^{n} \log \frac{1 - h_i}{1 - q_i} + \sum_{i=1}^{n} x_i \log \frac{h_i (1 - h_i)}{q_i (1 - q_i)} \geq 0 \\
&\Leftrightarrow u^T \begin{bmatrix} 1 \\ x \end{bmatrix} \geq 0
\end{aligned}
$$

where

$$
\begin{aligned}
u_0 &= \log \frac{p_1}{1 - p_1} + \sum_{i=1}^{n} \log \frac{1 - h_i}{1 - q_i} \\
u_i &= x_i \log \frac{h_i (1 - h_i)}{q_i (1 - q_i)}, \ i = 1, 2, ..., n.
\end{aligned}
$$