Machine Learning A (2023)
Home Assignment 4

Alexander Husted | wqg382

Contents

1 From a lower bound on the expectation to a lower bound on the probability

We want to prove the thighter bound $P(X \geq c) \geq \frac{a-c}{b-c}$
Let X be a random variable that is always upper bounded by b
Let $0 \leq x < a < b$
Let $X \in [0,1]$ and $c \in [0,1]$
Then by the inequality from Yevgeny's slides we have that:

$$E(Z) \leq c\mathbb{P}(Z \leq c) + 1\mathbb{P}(Z > c) \Rightarrow$$

$$a \leq c\mathbb{P}(X \leq c) + b\mathbb{P}(X > c) \Rightarrow$$

$$a \leq c(1 - \mathbb{P}(X \geq c)) + b\mathbb{P}(X > c) \Rightarrow$$

$$a \leq c - c\mathbb{P}(X \geq c) + b\mathbb{P}(X \geq c) \leq c - c\mathbb{P}(X \geq c) + b\mathbb{P}(x \geq c) \Rightarrow$$

$$a - c \leq -c\mathbb{P}(X \geq c) + b\mathbb{P}(X \geq c) \Rightarrow$$

$$a - c \leq (-c + b)\mathbb{P}(X \geq c) \Rightarrow$$

$$a - c \leq (b - c)\mathbb{P}(X \geq c) \Rightarrow$$

If X was not bounded the vaiable b would be infinitely large. Since b is the upper bound of X.

2 Learning by discretization

2.1 1

We define $\pi(h) = \frac{1}{M} = \frac{1}{|H_d|} = \frac{1}{2^{d^2}}$ (We have a d^2 grid where each square is either 1 or 0) such that we distribute the confidence budget δ uniformly among the hypotheses in $\mathcal{H}$ while satisfying $\pi(h) > 0 \forall h$ and $\sum_{h \in \mathcal{H}} \pi(h) \leq 1$. We derive the generalization bound from Occam's Razor Bound to distribute the uncertainty budget unevenly among the hypotheses in $\mathcal{H}_d$.

Theorem 3.3

$$P\left(\exists h \in \mathcal{H}_d : L(h) \geq \hat{L}(h,S) + \sqrt{\frac{ln\left(\frac{1}{\pi(h)\delta}\right)}{2n}}\right) \leq \delta \Rightarrow$$

$$P\left(\exists h \in \mathcal{H}_d : L(h) \geq \hat{L}(h,S) + \sqrt{\frac{ln\left(\frac{1}{\frac{1}{M}\delta}\right)}{2n}}\right) \leq \delta \Rightarrow$$

$$P\left(\exists h \in \mathcal{H}_d : L(h) \geq \hat{L}(h,S) + \sqrt{\frac{ln\left(\frac{M}{\delta}\right)}{2n}}\right) \leq \delta$$

$$P\left(\exists h \in \mathcal{H}_d : L(h) \geq \hat{L}(h,S) + \sqrt{\frac{ln\left(\frac{2^{d^2}}{\delta}\right)}{2n}}\right) \geq 1 - \delta$$

2.2 2

We choose

$$\pi(h) = \pi(H_d(h))\frac{1}{|H_d|} = \frac{1}{2^{d(h)+1}}\frac{1}{2^{d(h)^2}} = \frac{1}{2^{d(h)+1}}\frac{1}{2^{f(h)}} = \frac{1}{2^{d(h)+1+f(h)}}$$

The first part of $\pi(h)$ distributes the confidence budget δ among $\mathcal{H}_d - S$ and the second part distributes the confidence budget uniformly within $\mathcal{H}_d$

$$P\left(\exists h \in \mathcal{H} : L(h) \geq \hat{L}(h,S) + \sqrt{\frac{ln\left(\frac{1}{\frac{1}{2^{d(h)+1+f(h)}}\delta}\right)}{2n}}\right) \leq \delta \Rightarrow$$

$$P\left(\exists h \in \mathcal{H} : L(h) \geq \hat{L}(h,S) + \sqrt{\frac{ln\left(\frac{2^{d(h)+1+f(h)}}{\delta}\right)}{2n}}\right) \geq 1-\delta \Rightarrow$$

2.3 3

We want to find a d that minimizes

$$\hat{L}(h,S) + \sqrt{\frac{ln\left(\frac{2^{d(h)+1+f(h)}}{\delta}\right)}{2n}}$$

2.4 4

We have that the bound is true for all $h \in \mathcal{H}$, but it is only interesting when $d << log(n)$ But if we are interested in seeing how they scale comapred to each other, we can look at the most important faktors of the squareroot:

$$\frac{ln\left(\frac{2^{d(h)+1+f(h)}}{\delta}\right)}{2n} \approx \frac{2^d}{n}$$

2.5 5

We see that as $d(h)$ increasses, so does $\sqrt{\frac{ln\left(\frac{2^{d(h)+1+f(h)}}{\delta}\right)}{2n}}$ meaning that the when $d(h)$ incresses the bound becomes less tight.

3 Early Stopping

3.1 1

a) - Unbiased
This case is an unbiased estimate of $L(h_t*)$ since our desicions (wich model to pick and when to stop) is not decided based on the data.

b) - Biased
This case introduces bias, because we choose the model with the lowest validation error. And thus we make a decision based on S_{val}.

c) - Biased
This case introduces more bias than b, since we dicide both when to stop and which model to choose based on obersavations in the data.

3.2 2

a - Single Hypothesis)
We want to define a bound for a single hypothesis $h_{t^*} = h_{100}$ We use Theorem 3.1

$$P\left(L(h_{100}) \geq \hat{L}(h_{100}, S_{val}) + \sqrt{\frac{ln\frac{1}{\delta}}{2n}}\right) \geq 1 - \delta$$

b - Finite hypotheses spaces)
We want to define a bound for a finite number of hypthesis $h_1, h_2, ..., h_t$ such that we have $M = T$ hypothesis.
We use theorem 3.2

$$P\left(\exists h \in \mathcal{H} : L(h) \geq \hat{L}(h_{t^*}, S_{val}) + \sqrt{\frac{ln\frac{T}{\delta}}{2n}}\right) \geq 1 - \delta$$

c - Occam's Razor bound)
We use $\sum_{i=1}^{\infty} \frac{1}{i(i+1)} = 1$ to derive a $\pi(h)$ where $\sum_{h \in H} p(h) = 1$.

$$P\left(\exists h \in \mathcal{H} : L(h) \geq \hat{L}(h_{t^*}, S_{val}) + \sqrt{\frac{ln\left(\frac{1}{\pi(h)\delta}\right)}{2n}}\right) \leq \delta \Rightarrow$$

$$P\left(\exists h \in \mathcal{H} : L(h) \geq \hat{L}(h_{t^*}, S_{val}) + \sqrt{\frac{ln\left(\frac{1}{\frac{1}{T(T+1)}\delta}\right)}{2n}}\right) \leq \delta \Rightarrow$$

$$P\left(\exists h \in \mathcal{H} : L(h) \geq \hat{L}(h_{t^*}, S_{val}) + \sqrt{\frac{ln\left(\frac{T(T+1)}{\delta}\right)}{2n}}\right) \geq 1 - \delta$$

4 Logistic Regression

4.1 3.6

4.1.1 a

We want to show that

$$E_{in}(w) = \prod_{n=1}^{N} P(y_n|x_n) \Rightarrow$$

$$E_{in}(w) = \sum_{n=1}^{N}[y_n = +1]ln\frac{1}{h(x_n)} + \sum_{n=1}^{N}[y_n = -1]ln\frac{1}{1 - h(x_n)}$$

We know that:

$$P(y|x) = \begin{cases} h(x) & for\ \ y = +1 \\ 1 - h(x) & for\ \ y = -1 \end{cases}$$

Maximum likelyhood selects the hypothesis which maximizes the probability, but we can also minimize the negative of the probability.

$$-\frac{1}{N}ln\left(\prod_{n=1}^{N} P(y_n|x_n)\right) = \frac{1}{N}\sum^{N} ln\left(\frac{1}{P(y_n|x_n)}\right)$$

Since we have a Bernoulli distribution we can rewrite as two sums, one for each output of y.

$$-\sum_{n=1}^{N}[y_n = +1]ln(h(x_n)) + \sum_{n=1}^{N}[y_n = -1]ln(1 - h(x_n)) =$$

$$\sum_{n=1}^{N}[y_n = +1]ln\left(\frac{1}{h(x_n)}\right) + \sum_{n=1}^{N}[y_n = -1]ln\left(\frac{1}{1 - h(x_n)}\right) = E_{in}(w)$$

Where $[y_n = +1]$ is the probability that y is equal to 1.

4.1.2 b

We know that $\theta(s) = \theta\left(\frac{e^s}{1+e^s}\right)$

$$E_{in}(w) = \sum_{n=1}^{N} [y_n = +1] ln\left(\frac{1}{h(x_n)}\right) + \sum_{n=1}^{N} [y_n = -1] ln\left(\frac{1}{1-h(x_n)}\right) =$$

$$\sum_{n=1}^{N} [y_n = +1] ln\left(\frac{1}{\theta(w^T x)}\right) + \sum_{n=1}^{N} [y_n = -1] ln\left(\frac{1}{1-\theta(w^T x)}\right) =$$

$$\sum_{n=1}^{N} [y_n = +1] ln\left(\frac{1}{\frac{e^{w^T x}}{1+e^{w^T x}}}\right) + \sum_{n=1}^{N} [y_n = -1] ln\left(\frac{1}{1-\frac{e^{w^T x}}{1+e^{w^T x}}}\right) =$$

$$\sum_{n=1}^{N} [y_n = +1] ln\left(1 + e^{-w^T x_n}\right) + \sum_{n=1}^{N} [y_n = -1] ln\left(1 + e^{w^T x_n}\right)$$

Then we can combine $[y_n = 1]$ and $[y_n = -1]$ such that

$$E_{in}(w) = \sum_{n=1}^{N} ln(1 + e^{-y_n w^T x_n})$$

4.2 3.7

We know that

$$\nabla E_{in}(w) = -\frac{1}{N}\sum_{n=1}^{N}\frac{y_n x_n}{1+e^{y_n w^T x_n}}$$

Which is equal to the derivative of $E_{in}(w)$ with respect to w. And we know that $\left(\frac{e^s}{1+e^s}\right) = \theta(s)$

$$\frac{d}{dw}\left(-\frac{1}{N}\sum_{n=1}^{N} ln(1+e^{-y_n w^T x_n}\right) =$$

$$-\frac{1}{N}\sum_{n=1}^{N}\frac{1}{1+e^{-y_n w^T x_n}}\frac{d}{dw}\left(1+e^{-y_n w^T x_n}\right) =$$

$$-\frac{1}{N}\sum_{n=1}^{N} y_n x_n \frac{e^{-y_n w^T x_n}}{1+e^{-y_n w^T x_n}} =$$

$$-\frac{1}{N}\sum_{n=1}^{N} y_n x_n \theta(-y_n w^T x_n) =$$

$$\frac{1}{N}\sum_{n=1}^{N} -y_n x_n \theta(-y_n w^T x_n)$$

A point is classified correctly if $y_n w^T x_n > 0$ and misclassified if $y_n w^T x_n \leq 0$
We have that

$$y_n w^T x_n \leq 0 \Rightarrow \frac{e^{-y_n w^T x_n}}{1+e^{-y_n w^T x_n}} \geq 0.5$$

This means that when we take the mean $\frac{1}{N}\sum_{n=1}^{N}$ the misclassified contributes more then the correctly classified ones.

The {0,1} case:

One could argue that

$$\left[\frac{y_n+1}{2} - \theta(w^T x_n)\right] x_n = \begin{cases} \left[1-\theta(w^T x_n)\right] x_n & For\ y_n = 1 \\ \left[0-\theta(w^T x_n)\right] x_n & For\ y_n = -1 \end{cases}$$

Which in the {0,1} case is equvilant to

$$\left[y_n - \theta(w^T x_n)\right] x_n = \begin{cases} \left[1-\theta(w^T x_n)\right] x_n & For\ y_n = 1 \\ \left[0-\theta(w^T x_n)\right] x_n & For\ y_n = 0 \end{cases}$$

This means that the argument in the second case is the same as the first.

4.3 Log-odds

Let $s = w^T x + b$ and let $P(Y = 1|X = x) = y$ we assume that s encodes the log-odds:

$$s = ln\left(\frac{P(Y = 1|X = x)}{P(Y = 0|X = x)}\right) = ln\left(\frac{y}{1-y}\right)$$

If σ is the logistic function we shoud have that $\sigma(s) = y$ (Keep in mind the fraction rule $\frac{\frac{a}{b}}{c} = \frac{a}{bc}$)

$$\sigma(s) = \frac{e^s}{1+e^s} = \frac{e^{ln\left(\frac{y}{1-y}\right)}}{1+e^{ln\left(\frac{y}{1-y}\right)}} = \frac{\frac{y}{1-y}}{1+\frac{y}{1-y}} = \frac{y}{(1-y)(1+\frac{y}{1-y})} = y$$

Thus we have that if the (affine) linear part of the model ($w^T x + b$) encodes the log-odds then σ is the logistic function