Machine Learning A (2023)
Home Assignment 4

Niels Krarup WTG176

Contents

1 From a lower bound on the expectation to a lower bound on the probability (5 points)

Recall that for a continuous random variable $Z \in [0,1]$ we have $\mathbb{E}Z \leq c\mathbb{P}(Z \leq c)+\mathbb{P}(Z \geq c)$ for $c \in (0,1)$. This follows from splitting up the integral above and below c and replacing the integrand with c when we integrate over the area where $(Z \leq c)$, and 1 when $(Z > c)$

Since X is bounded by $b > 0$ we can use the above on X/b and use Markov to write:

$$\begin{aligned}
&\mathbb{E}\frac{X}{b} \leq \frac{c}{b}\mathbb{P}\left(\frac{X}{b} \leq \frac{c}{b}\right) + \mathbb{P}\left(\frac{X}{b} \geq \frac{c}{b}\right) \Leftrightarrow \\
&\frac{1}{b}\mathbb{E}X \leq \frac{c}{b}\mathbb{P}\left(X \leq c\right) + \mathbb{P}\left(X \geq c\right) \Leftrightarrow \\
&\mathbb{P}\left(X \geq c\right) \geq \frac{1}{b}a - \frac{c}{b}\mathbb{P}\left(X \leq c\right) \Leftrightarrow \\
&\mathbb{P}\left(X \geq c\right) \geq \frac{a-c}{b}
\end{aligned}$$

Where we have used that $\mathbb{E}X = a$, that $b > 0$ such that we can multiply on both sides of the inequalities without them flipping, that $\mathbb{P}(.) \leq 1$ and linearity of the mean.

2 Learning by discretization (20 points)

2.1

From the construction of $\mathcal{H}_d$ we see a total of $d \times d$ uniform squared with binary decisions. Hence a total of 2^{d^2} hypothesis in $\mathcal{H}_d$. Hence the $|\mathcal{H}_d| = 2^{d^2} = M$ such that we can use THM 3.2 to get a generalization bound for learning with $\mathcal{H}_d$:

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

2.2

If we now want to use $\mathcal{H} = \bigcup_{d=1}^{\infty}$ for learning, the size in infinite so we need to use weights on each hypothesis set $\pi(d)$ to balance the increased number of hypothesis in each $\mathcal{H}_d$ with the increase in precision i.e. performance and complexity.

The general form of π will be

$$\pi(h) = p\left(\mathcal{H}_{d(h)}\right) \times \frac{1}{2^{d(h)^2}}$$

Where the second term distributes the confidence budget uniformly within $h \in \mathcal{H}_d$. Now, the first term are weights given to each class, which simply has to sum to 1. There are many choices, but for simplicity and to mimic THM 3.5 we choose $p(h) = \frac{1}{2^{d(h)}}$, which sums to one since $d(h) \in \{1, 2, \dots\}$ and that is is seen to be a geometric series with $r = \frac{1}{2} < 1$. Then, following THM 3.3 we get:

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

2.3

As stated earlier the weighting of $\mathcal{H}_d$ is arbitrary and any discrete distribution will work, as any distribution by definition sums to 1. One can use the shape of the distribution to select for complexity vs simplicity. If there is prior information that breaks the permutation symmetry it can be used to assign higher prior to the corresponding trees and if it correctly reflects the true data distribution it will also lead to tighter bounds.

2.4

2.5

3 Early Stopping (30 points)

3.1

3.1.a

3.1.b

3.1.c

3.2

3.3

3.4

3.5

3.5.a

3.5.b [optional]

3.5.c

3.5.d [optional]

4 Logistic Regression

4.1 Cross-entropy error measure (15 points)

a) One of the mathematical convenient aspects of using the logistic function, $\theta(s) = \frac{1}{1+e^{-s}}$, as model is that the likelihood of a data point (y_n, x_n) can be written compactly as $\mathbb{P}(y_n|x_n) = \theta(y_n w^T x_n)$, due to the fact that $\theta(-s) = 1 - \theta(s)$. When this symetry is not the case for general models h we need to split up the likelihood into cases where $(y_n = +1)$ in which case the probability is $h(x_n)$ repspectivly the cases $(y_n = -1)$ which should then have probability $1 - h(x_n)$

In the full likelihood this can be written compactly by using the indicators $\mathbb{1}(y_n = \pm 1)$ as exponents, as follows:

$$L(y|x) = \prod_{n=1}^{N} h(x_n)^{\mathbb{1}(y_n=1)}(1 - h(x_n)^{\mathbb{1}(y_n=-1)}$$

of course the h that maximizes this quantity will analogously minimize minus the log of the above:

$$\begin{aligned}
-\log L(y|x) &= -\log \prod_{n=1}^{N} h(x_n)^{\mathbb{1}(y_n=1)}(1-h(x_n)^{\mathbb{1}(y_n=-1)} \\
&= -\sum_{n=1}^{N} \log \left\{ h(x_n)^{\mathbb{1}(y_n=1)}(1-h(x_n))^{\mathbb{1}(y_n=-1)} \right\} \\
&= \sum_{n=1}^{N} \mathbb{1}(y_n=1) \log \frac{1}{h(x_n)} + \mathbb{1}(y_n=-1) \log \frac{1}{1-h(x_n)}
\end{aligned}$$

b)

Let's start by confirming the fact that for θ the logistic function then $1-\theta(s)=\theta(-s)$:

$$\theta(-s) = \frac{1}{1+e^s} = \frac{e^{-s}}{e^{-s}+1} = \frac{1+e^{-s}-1}{e^{-s}+1} = 1-\theta(s)$$

This allows us to write the in-sample error E_{in} from a) as:

$$\begin{aligned}
&\sum_{n=1}^{N} \mathbb{1}(y_n=1) \log \frac{1}{\theta(w^T x_n)} + \mathbb{1}(y_n=-1) \log \frac{1}{1-\theta(w^T x_n)} = \\
&\sum_{n=1}^{N} \mathbb{1}(y_n=1) \log \frac{1}{\theta(w^T x_n)} + \mathbb{1}(y_n=-1) \log \frac{1}{\theta(-w^T x_n)} = \\
&\sum_{n=1}^{N} \mathbb{1}(y_n=1) \log \left(1+e^{-w^T x_n}\right) + \mathbb{1}(y_n=-1) \log \left(1+e^{w^T x_n}\right)
\end{aligned}$$

now, we see that in the case $y_n=1$ we use the left side of the sum with the exponent minus the linear predictor, and in the case $y_n=-1$ we use the right side of the sum with exponent equal to the linear predictor, in any case inserting y_n in front of $w^T x_n$, will give us exactly the part we use in any case.

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

of course multiplying by $\frac{1}{N}$ does not change the minimizing objective and we see the above as equivalent to (3.9)

4.2 Logistic regression loss gradient (15 points)

We simply take the partial derivative of the in sample error / likelihood function with respect to the element weight w_i and get:

$$\begin{aligned}\frac{\partial}{\partial w_i}E_{in}(w) &= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{1+e^{-y_n w^T x_n}}(-y_n x_{ni})e^{-y_n w^T x_n}\\ &= -\frac{1}{N}\sum_{n=1}^{N}\frac{y_n x_{ni}}{1+e^{y_n w^T x_n}}\end{aligned}$$

Where we have used the chain rule. Since this is element wise, we see that the whole gradient can be compactly written as

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

A misclassified data point, would be a $y_n = 1(-1)$ which has associated predictor $w^T x_n$ small (large). On the other hand a correctly specified sample would agree between y_n and $w^T x_n$ in the sense that $y_n = 1$ corresponds to large $w^T x_n$ and vice versa. In that case the denominator would be small and hence the change in w would not effect a lot in the gradient. On the other hand for a misclassified sample the denominator is not large and hence w has a bigger effect.

If we use $y_n \in \{0, 1\}$ instead of ± 1 we get the slightly altered likelihood/ in-sample error, as the last step in b) above would become:

$$\begin{aligned}-\log \tilde{L}(y|x) &= \sum_{n=1}^{N}\mathbb{1}(y_n = 1)\log\left(1+e^{-w^T x_n}\right) + \mathbb{1}(y_n = 0)\log\left(1+e^{w^T x_n}\right)\\ &= \sum_{n=1}^{N}\log\left(1+e^{(1-2y)w^T x_n}\right)\end{aligned}$$

following the lines from before we start by finding the partial derivative:

$$\begin{aligned}\frac{\partial}{\partial w_i}\tilde{E}_{in}(w) &= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{1+e^{(1-2y_n)w^T x_n}}(1-2y_n)x_{ni}e^{(1-2y_n)w^T x_n}\\ &= \frac{1}{N}\sum_{n=1}^{N}\theta\left((1-2y_n)w^T x_n\right)(1-2y_n)x_{ni}\\ &= -\frac{1}{N}\sum_{n=1}^{N}\left(y-\theta(w^T x_n)\right)x_{ni}\end{aligned}$$

Where we have used the fact that $1-\theta(s) = \theta(-s)$. Combining this one gets the compact vector formula:

$$\nabla\tilde{E}_{in}(w) = -\frac{1}{N}\sum_{n=1}^{N}\left(y-\theta(w^T x_n)\right)x_n$$

again, when $y = 1$ then $w^T x$ should be large for a correctly classification which makes the term in the sum small, in $y = 0$ and $w^T x$ large then the sum will be close to -1 and the w will influence more, and vice versa.

4.3 Log-odds (15 points)

We simply plug in the assumption that the linear predictor, $\omega^T x + b$ equals the log-odds for σ being the logistic function $\sigma(u) = \frac{1}{1+e^{-u}}$.

$$\begin{aligned} f(x) &= \sigma(\omega^T x + b) \\ &= \frac{1}{1 + e^{-\log \frac{\mathbb{P}(Y=1|x)}{\mathbb{P}(Y=0|x)}}} \\ &= \frac{1}{1 + \frac{\mathbb{P}(Y=0|x)}{\mathbb{P}(Y=1|x)}} \\ &= \frac{\mathbb{P}(Y = 1|x)}{\mathbb{P}(Y = 1|x) + \mathbb{P}(Y = 0|x)} \\ &= \mathbb{P}(Y = 1|x) \end{aligned}$$

Now of course, we have shown that the logistic function does yield the probability of $(Y = 1)$ if the linear predictor encodes the log odds. But there of course may be other functions $\tilde{\sigma}$ which does the same, but are not the logistic function. However, since the above holds for all values of the linear predictor $\omega^T x + b \in \mathbb{R}$ we must have that

$$\sigma(u) = \tilde{\sigma}(u),\ u \in \mathbb{R}$$

Hence all other functions will be on the same form i.e. logistic function.