Machine Learning A - Home Assignment 3

Jakob H. Schauser, pwn274

September 2021

1 Distribution of Student's Grades

My god did I spend a lot of time trying to make sense of the answers in this problem - only to realize after talking to the others, that they are not supposed to make sense.

1.1 Markov

Taking the hint $\rightarrow$ Optimizing for Q is the same as minimum for Z

$$\mathbb{P}(\hat{Q} \geq \varepsilon) = \mathbb{P}(100 - \hat{Q} \leq 100 - \varepsilon) = \mathbb{P}(\hat{Z} \leq 100 - \varepsilon) \leq \frac{\mathbb{E}[\hat{Q}]}{\varepsilon} = \delta \tag{1}$$

Now defining $z = 100 - \varepsilon$

$$\delta = \frac{\mathbb{E}[\hat{Q}]}{100 - z} = \frac{\mathbb{E}[100 - \hat{Z}]}{100 - z} = \frac{100 - \mathbb{E}[\hat{Z}]}{100 - z} = \frac{100 - p}{100 - z} = \frac{100 - 60}{100 - z} = 0.05 \tag{2}$$

$$z = 100 - \frac{40}{0.05} = -700 \tag{3}$$

Nonsense!

1.2 Chebyshev

Only looking the one way must be less likely.

$$\mathbb{P}(|\hat{Z} - p| \geq \epsilon) \geq \mathbb{P}(\hat{Z} \leq p - \epsilon) \tag{4}$$

Now using Chebyshev $Y = 0.4 * 0 + 0.6 * 100$

$$\mathbb{P}(\hat{Z} \leq p - \epsilon) \leq \frac{Var[Z]}{\epsilon^2} \leq \frac{Var[Y]}{\epsilon^2} = \frac{\mathbb{E}[Y^2] - \mathbb{E}[Y]^2}{\epsilon^2} = \frac{0.6 * 100^2 - p^2}{\epsilon^2} = \delta \tag{5}$$

Using $z = p - \epsilon$

$$\mathbb{P}(\hat{Z} \leq z) \leq \frac{2400}{(p - z)^2} = \delta \leftrightarrow z = -159 \ \ \& \ \ 279 \tag{6}$$

Nonsense!

1.3 Hoeffding

Dividing Z by 100 to get it into the [0,1]-range to use corollary 2.5

$$\mathbb{P}\left(\mu - \frac{1}{n}\sum_{i=1}^{n} Z_i \geq \varepsilon\right) = \mathbb{P}\left(p - \hat{Z} \geq \varepsilon\right) = \mathbb{P}\left(\hat{Z} \leq p - \varepsilon\right) \leq e^{-2n\varepsilon^2} = \delta \tag{7}$$

Using $z = p - \varepsilon$

$$e^{-2n(p-z)^2} = \delta \leftrightarrow z' = 0.137 \ \ \& \ \ 1.06 \tag{8}$$

Giving us the first sane answer of about z = 13.7

1.4 Which of the three inequalities provide a non-vacuous value of z?

Only Hoeffding as the others both gave a results less than 0 and above 100, which are trivially right and trivially wrong respectively.

2 How to Split a Sample into Training and Test Sets

2.1

I start out by the ordinary Hoeffding inequality:

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \geq \varepsilon'\right) \leq e^{-2\varepsilon'^2/\sum_{i=1}^{n}(b_i-a_i)^2} \tag{9}$$

Dividing through with n and using $n = \varepsilon'/n$

$$\mathbb{P}\left(1/n\sum_{i=1}^{n} X_i - 1/n\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \geq \varepsilon\right) = \mathbb{P}\left(L - \hat{L} \geq \varepsilon\right) \leq e^{-2(n\varepsilon)^2/\sum_{i=1}^{n}(b_i-a_i)^2} \tag{10}$$

I will now isolate the bound, ε:

$$e^{-2(n\varepsilon)^2/\sum_{i=1}^{n}(b_i-a_i)^2} = \delta \leftrightarrow \varepsilon = \sqrt{\frac{\ln(\delta)\sum_{i=1}^{n}(b_i-a_i)^2}{-2n^2}} = \sqrt{\frac{\ln(1/\delta)\sum_{i=1}^{n}(b_i-a_i)^2}{2n^2}} \tag{11}$$

Doing some moving around and inserting ε:

$$\mathbb{P}\left(L - \hat{L} \geq \varepsilon\right) \leq \delta \leftrightarrow \mathbb{P}\left(\hat{L} \geq L - \sqrt{\frac{\ln(1/\delta)\sum_{i=1}^{n}(b_i-a_i)^2}{2n^2}}\right) \geq 1 - \delta \tag{12}$$

In words, this can be read as $\hat{L}$ being bounded by $L - \sqrt{\frac{\ln(1/\delta)\sum_{i=1}^{n}(b_i - a_i)^2}{2n^2}}$ with a probability of at least $\geq 1 - \delta$.

This of course simplifies down to the the well known result from corollary 2.5 if we can choose the bounds of our loss function freely.

2.2

Ran out of time - I am looking forward to seeing the solution to this problem, as working with it, it seemed highly non-trivial.

3 Linear classification

3.1 Cross-entropy error measure

a)

Here most of the legwork is already done in the book. It is shown that for a hypothesis function following:

$$P(y \mid \mathbf{x}) = \begin{cases} h(\mathbf{x}) & \text{for } y = +1 \\ 1 - h(\mathbf{x}) & \text{for } y = -1 \end{cases} \tag{13}$$

We have that the negative log-likelihood is given by:

$$\text{NLLH} = \frac{1}{N}\sum_{n=1}^{N} \ln\left(\frac{1}{P\left(y_n \mid \mathbf{x}_n\right)}\right) \tag{14}$$

Now given that any factor of 1/N doesn't matter because of argmax we can remove this. We also know every case and the probability of each, allowing us to separate them:

$$\text{NLLH} = \sum_{n=1}^{N} [[y_n = +1]] \ln\left(\frac{1}{h(\mathbf{x}_n)}\right) + [[y_n = -1]] \ln\left(\frac{1}{1 - h(\mathbf{x}_n)}\right) \tag{15}$$

b)

Using the fact that $\theta(-x) = 1 - \theta(x)$

$$
\begin{aligned}
E_{in} &= \sum_{n=1}^{N} [[y_n = 1]] \ln \frac{1}{\theta(\boldsymbol{\omega}^\top \mathbf{x_n})} + [[y_n = -1]] \ln \frac{1}{1 - \theta(\boldsymbol{\omega}^\top \mathbf{x_n})} \\
&= \sum_{n=1}^{N} [[y_n = 1]] \ln \frac{1}{\theta(\boldsymbol{\omega}^\top \mathbf{x_n})} + [[y_n = -1]] \ln \frac{1}{\theta(-\boldsymbol{\omega}^\top \mathbf{x_n})} \\
&= \sum_{n=1}^{N} \ln \frac{1}{\theta(y_n \boldsymbol{\omega}^\top \mathbf{x_n})} = \sum_{n=1}^{N} \ln(1 + e^{y_n \boldsymbol{\omega}^\top \mathbf{x_n}})
\end{aligned}
$$

Which is exactly what we wanted to end up with.

QED

3.2 Logistic regression loss gradient

For logistic regression the in-sample-error is:

$$
E_{\text{in}}(\boldsymbol{\omega}) = \frac{1}{N} \sum_{n=1}^{N} \ln \left(1 + e^{-y_n \boldsymbol{\omega}^\top \mathbf{x}_n}\right) \tag{16}
$$

I use the fact that $\theta = \frac{1}{1+e^{-x}}$

$$
\begin{aligned}
\nabla E_{\text{in}} &= \frac{1}{N} \sum \frac{1}{1 + e^{-y_n \boldsymbol{\omega}^\top \mathbf{x}_n}} \cdot (-y_n \mathbf{x}_n) e^{-y_n \boldsymbol{\omega}^\top \mathbf{x}_n} \\
&= \frac{-1}{N} \sum \frac{1}{1 + e^{y_n \boldsymbol{\omega}^\top \mathbf{x}_n}} \cdot (y_n \mathbf{x}_n) \\
&= \frac{1}{N} \sum -y_n \mathbf{x}_n \cdot \theta(y_n \boldsymbol{\omega}^\top \mathbf{x}_n)
\end{aligned}
$$

3.3 Log-odds

Here I simply insert and reduce, showing that they are equal:

$$
\begin{aligned}
\boldsymbol{\omega}^\top \mathbf{x} + b &= \ln \frac{P(Y = 1 \mid X = x)}{P(Y = 0 \mid X = x)} = \ln \frac{P(Y = 1 \mid X = x)}{1 - P(Y = 1 \mid X = x)} \\
&= \ln \frac{P(Y = 1 \mid X = x)}{1 - P(Y = 1 \mid X = x)} = \ln \frac{\sigma(\boldsymbol{\omega}^\top \mathbf{x} + b)}{1 - \sigma(\boldsymbol{\omega}^\top \mathbf{x} + b)} = \ln \frac{1}{1/\sigma(\boldsymbol{\omega}^\top \mathbf{x} + b)} \\
&= \ln \frac{1}{(1 + e^{-\boldsymbol{\omega}^\top \mathbf{x} + b}) - 1} = \ln \frac{1}{e^{-(\boldsymbol{\omega}^\top \mathbf{x} + b)}} = \ln e^{\boldsymbol{\omega}^\top \mathbf{x} + b} = \boldsymbol{\omega}^\top \mathbf{x} + b
\end{aligned}
$$

3.4 Variable importance

There is some inherent redundancy in using one-hot encoding. Much like fitting, unnecessary degrees of freedom are unwanted. As any single class simply can be expressed in

terms of the others, one of the weights trained to predict a given class is unnecessary. This is of course bad for computational power, but also has consequences, as having one free variable suddenly gives infinite 'equal' solutions to every regression.

The problem of interpretability I have a harder time imagining, but I think, the free variable gives one of the predicted probabilities too much free room. This could, among other things, lead to the predicted probabilities not summing to 1, which would be optimal if only the important values were regressed.