Assignment 2

Jeppe Johansen

December 5, 2017

1 Illustration of Hoeffding's Inequality

1

The implementation of the coin flips is made the following way. Using a binomial distribution (repeated bernoulli distribution), with $\mu = 0.5, n = 20$ times 1000000 I get the right distribution. This can be considered a 20×1000000 matrix.

2

Below is the plot of the frequencies. of getting $P(\frac{1}{N}\sum_{i=1}^{N} X \geq \alpha), \alpha \in [0.5, 0.55, \ldots, 0.95, 1]$ and $N = 20$. As the figure shows the probability decreases as the right hand side of a normal distribution. This is not a surprise however, since the central limit theorem ensures that the parameter space θ is normally distributed. Here it's noted that θ is the empirical average.

It's noted

3

It's not necessary to add any more granularity to the distribution of alpha, since the empirical average og 20 coin flips cannot take any discrete value inbetween the given alphas. This follows clearly from this example: Assume 20 coin flips with 10 heads and 10 tails. In this case the empirical average would be 0.5. Now assume that in the next 20 coin flips there will be 11 heads and 9 tails. This would lead to the empirical average of 0.55, which imply a granularity of 0.05

6

Looking at figure 2 Hoeffding's bound and markov's bound is plotted next to the values of frequencies of observed empirical averages above 5. First it's noted that the sharpest decline happens in the frequencies of the empirical averages. This is totally expected, since both Markov's and Hoeffding's bound should be an upper limit of a given observed frequency of the empirical average. Next it's

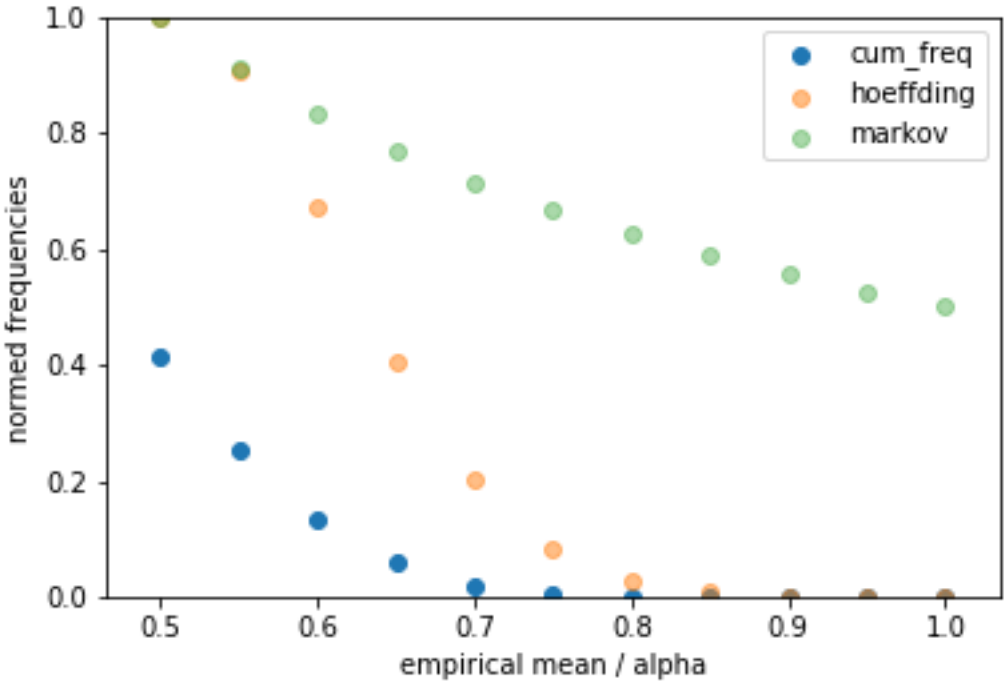

Figure 1: Distributions of Z

noted that hoeffdings bound declines much faster that the Markov bound. This has the implication, that Hoeffding's bound is a better way to bound probability that Markov's bound, since, when bounding probability it's prefered that have as hard a bound as possible, to get a sense of the possible outcomes of a given random process. Lastly it can be seen that Markov's bound decrease way slower than Hoeffding's bound, however it still does bound the probability space.

7

The exact probability is calculated the following way $N = 20$:

$$\mathbb{P}\left\{\frac{1}{N}\sum_{i=1}^{N} X_i \geq= 1\right\} = \prod_{i=1}^{N}\frac{1}{2} = 9.537 \times 10^{-7}$$

The probability that the empirical average is larger or equal to 0.95

$$\mathbb{P}\left\{\frac{1}{N}\sum_{i=1}^{N} X_i \geq= 0.95\right\} = \sum_{\alpha\in\{0.95,1\}} \mathbb{P}(X \geq \alpha) = 2.003 \times 10^{-5}$$

The hoefding's bounds HB is calculated below:

$$HB(\alpha = 1.00, \mu = 0.5, N = 20) = 4.540 \times 10^{-5}$$
$$HB(\alpha = 0.95, \mu = 0.5, N = 20) = 0.0003035$$

The Markov's bounds MB is calculated below:

$$MB(\alpha = 1.00, \mu = 0.5, N = 20) = 0.5$$
$$MB(\alpha = 1.00, \mu = 0.5, N = 20) = 0.526$$

2 The effect of scale (range) and normalization of random variables in Hoeffding?s inequality

From the Theorem of Hoeffding's Inequality I get:

$$\mathbb{P}\left\{\sum_{i}^{N} X_i - \sum_{i}^{N}\mathbb{E}\left[X_i\right] \geq \epsilon\right\} \leq e^{-2\epsilon/\sum_{i}^{N}(b_i-a_i)^2} \tag{1}$$

Since I know we are looking at discrete data (people either arriving or not I know: $\mathbb{P}\{X_i \in [0,1]\} = 1$ and $\mathbb{E}[X_i] = \mu$ for all observations. This implies:

$$a_i = 0 \wedge b_i = 1 \tag{2}$$

Using this I rewrite:

$$\mathbb{P}\left\{\sum_{i}^{N} X_i - \sum_{i}^{N} \mathbb{E}\left[X_i\right] \geq \epsilon\right\} \leq e^{-2\epsilon/N} \tag{3}$$

DIviding through by n yields:

$$\mathbb{P}\left\{\frac{1}{N}\sum_{i=1}^{N} X_i - \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[X_i\right] \geq \frac{1}{N}\epsilon\right\} \leq e^{-2\epsilon/N} \tag{4}$$

$$= \quad \mathbb{P}\left\{\frac{1}{N}\sum_{i=1}^{N} X_i - \mathbb{E}\left[X_i\right] \geq \frac{1}{N}\epsilon\right\} \leq e^{-2\epsilon/N} \tag{5}$$

Now it's clear that epsilon ca be redefined: $\hat{\epsilon} = \epsilon\frac{1}{N}$

$$\mathbb{P}\left\{\frac{1}{N}\sum_{i=1}^{N} X_i - \mathbb{E}\left[X_i\right] \geq \hat{\epsilon}\right\} \leq e^{-2n\hat{\epsilon}} \tag{6}$$

Which was the result wanted (corollary 2.4)

(7)

3 Probability in Practice

3.1 1

An airline has 99 seats in the plane. We know they sell 100 tickets. The probability of not showing up i equal 0.05 which implies $\mathbb{P}[Showingup] = 0.95$ since there is only 1 way that over 99 people can board the plane we can calculate the bounds.

$$\mathbb{P}\left\{X > 99\right\} = \mathbb{P}\left\{X \geq 100\right\} = \prod_{i=1}^{100} 0.95 = 0.00592 \tag{8}$$

Furthermore the Hoeefding's bound is found:

$$\mathbb{P}\left\{\frac{1}{N}\sum_{i=1}^{N} \geq \underbrace{\alpha - \mu}_{\epsilon} = 1 - 0.95\right\} \leq e^{2n\epsilon^2} = 0.6065 \qquad N = 100 \tag{9}$$

Lastly the Markov's bound is found

$$\mathbb{P}\left\{\frac{1}{N}\sum_{i=1}^{N} \geq \epsilon = 1\right\} \leq \frac{\mathbb{E}[X]}{\epsilon} = \frac{0.95}{1} \qquad N = 100 \tag{10}$$

3.2 2

We have to different set of information given. That empirically we found that 9500 out of 10000 guests arrive, and we know that we have overbooked if all passengers arrive to a given plane. We have from 3.1 already found that the probability of a plane being overbooked can be described as p^{100} where 100 denote the number of passengers. I optimize the function by knowing that these to probabilities are independent that is the probability of observing 9500 out of 10000 people getting to their plane and all 100 arriving to one giving plane. Using this information we use the hoeffding bound for the probability of 9500 arriving to a plane. When this bound is found i multiply it with the probability that i a plane gets overbooked. Then i solve for p and find the bound.

$$\mathbb{P}\left(\sum_{i=1}^{10000} X_i \geq 9500\right) \tag{11}$$

$$=\mathbb{P}\left(\sum_{i=1}^{10000} X_i - 10000 \cdot p \geq 9500 - 10000 \cdot p\right) \tag{12}$$

$$\leq e^{-2(10000(0.95-p))^2/\sum_{i=1}^{10000}(b_i-a_i)^2} \tag{13}$$

$$= \leq e^{-2(10000(0.95-p))^2/\sum_{i=1}^{10000}(0-1)^2} \tag{14}$$

$$= \leq e^{-2(10000(0.95-p))^2/\sum_{i=1}^{10000}1} \tag{15}$$

$$= \leq e^{-2(10000(0.95-p))^2/10000} \tag{16}$$

Now using the expression for the probability (bound) of all people arriving to the plane p^{100} the probability can be bound:

$$p^{100} \cdot e^{-2(10000(0.95-p))^2/\sum_{i=1}^{10000}1} = 0.00619 \tag{17}$$

Where the above equation is solved numerically. We find $p = 0.954$ when the equation is solved numerically. The figure below (figure 2 shows the corresponding bound given probability p.

4 Logistic Regression

4.1 Cross entropy error measure

$$min. \qquad \frac{1}{N}\sum_{i=1}^{N} ln\left(\frac{1}{\theta(y_n \mathbf{w}^t \mathbf{x}_n)}\right) \qquad = \tag{18}$$

Using that $1 - \theta(s) = \theta(-s)$ I write:

$$\theta(y, \mathbf{x}, \mathbf{w}) = \mathbb{1}_{(y=1)}\theta(\mathbf{w}^t\mathbf{x}_n) + \mathbb{1}_{(y=-1)}\theta(-\mathbf{w}^t\mathbf{x}_n) \tag{19}$$

$$\theta(y, \mathbf{x}, \mathbf{w}) = \mathbb{1}_{(y=1)}\theta(\mathbf{w}^t\mathbf{x}_n) + \mathbb{1}_{(y=-1)}(1 - \theta(\mathbf{w}^t\mathbf{x}_n)) \tag{20}$$

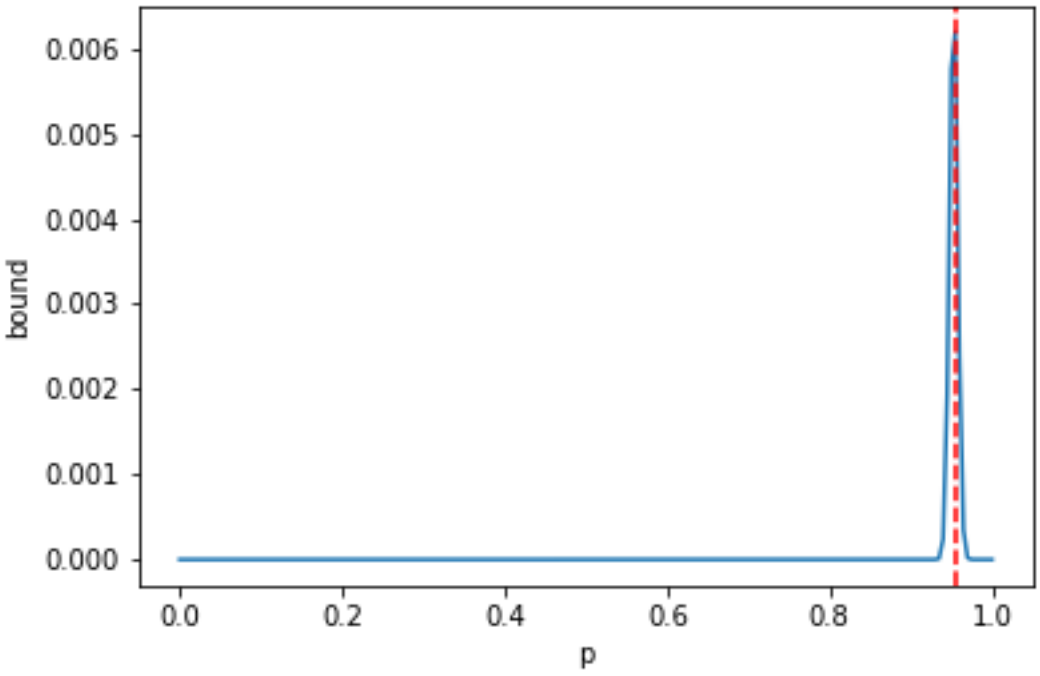

Figure 2: p and bound

Substituting $\theta(y, \mathbf{w}, \mathbf{x})$ with $h(x)$:

$$= \mathbb{1}_{(y=1)} \frac{1}{h(x)} + \mathbb{1}_{(y=-1)} \frac{1}{1-h(x)} \tag{21}$$

now plugging this result in:

$$min. \qquad \frac{1}{N} \sum_{i=1}^{N} \left(\mathbb{1}_{(y=1)} ln \frac{1}{h(x)} + \mathbb{1}_{(y=-1)} ln \frac{1}{1-h(x)} \right) \tag{22}$$

$$E_{(\mathbf{w})} = \frac{1}{N} \sum_{i=1}^{N} \left(\mathbb{1}_{(y=1)} ln \frac{1}{h(x)} + \mathbb{1}_{(y=-1)} ln \frac{1}{1-h(x)} \right) \tag{23}$$

4.2 Logistic regression loss gradient

Using the expression for $E_{in}(\mathbf{w})$ at equation 3.9:

$$E_{in}(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^{N} ln(1 + e^{-y_n \mathbf{w}^t \mathbf{x}_n}) \qquad \Leftrightarrow \tag{24}$$

$$\frac{\partial}{\partial \mathbf{w}} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{1 + e^{-y_n \mathbf{w}^t \mathbf{x}_n}} e^{-y_n \mathbf{w}^t \mathbf{x}_n} (-y_n \mathbf{x}_n) \tag{25}$$

$$\frac{\partial}{\partial \mathbf{w}} = \nabla E_{in}(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^{N} \frac{(-y_n \mathbf{x}_n)}{1 + e^{y_n \mathbf{w}^t \mathbf{x}_n}} \tag{26}$$

First noting that:

$$\theta(-s) = 1 - \theta(s) = 1 - \frac{e^s}{1+e^s} = \frac{e^s + 1 - e^s}{1+e^s} = \frac{1}{1+e^s} \tag{27}$$

using this expression we get:

$$\nabla E_{in}(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{1 + e^{y_n \mathbf{w}^t \mathbf{x}_n}} (-y_n \mathbf{x}_n) \tag{28}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \theta(y, \mathbf{w}, \mathbf{x}) (-y_n \mathbf{x}_n) \tag{29}$$

$$\tag{30}$$

4.3 Logistic regression implementation

The implementation of the logistic regression is made in the following way. Looking at $p.95$ the algorithm for the logistic regression is written up. First i calculate the scalar using matrix manipulation. From there i use a while loop that runs until the gradient is sufficiently small, which implies the parameters has converged sufficiently. When parameters has converged the the unique minimum of the parameter space has been found. The loop goes the following way:

1. Initialize weights $\vec{0}$ to the zero-vector.
2. calculate the gradient as described above.
3. recalculate the weights.
4. feed the new weights into the algorithm

4.4 Iris Flower Datset

In the iris flower dataset I implement the algorithm and get the following weigths in the test dataset:

1. Length: 8.48
2. Width: -50.84
3. Intercept: -29.04

And the test error is: $\frac{1}{N}\sum_{i=1}^{N} error = 0.61$ where error is $error \in \{0, 1\}$ This is a quite poor performance, and could imply an error in the code. However after two different implementations of both the gradient-algorithm and the loop-algorithm, i must concede and conclude the algorithm just performs poorly.