Machine Learning
Assignment 1.2

Nikolaj Dybdahl Rathcke (rfq695)

November 29, 2015

1 Linear Regression

1.1

The algorithm is implemented in the file `linear_regr.py` as the function `linReg`. The function takes one argument, a matrix, and returns the matrix $[w, b]$ - the parameters for a linear equation $y = wx + b$.

1.2

The results from running produces the two parameters are

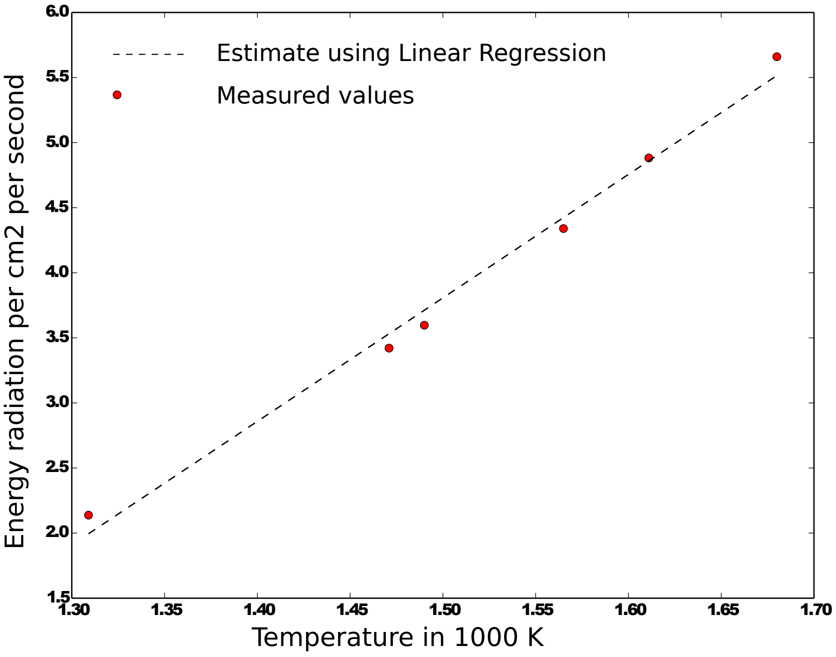
```
Parameters:
[  9.48934569 -10.42696146]

Mean Squared Error:
0.0124342216151
```

where $w \approx 9.49$ and $b \approx -10.43$.

1.3

The plot is obtained by using the library `matplotlib` which produces the following figure:



where the line is the estimation for y and the red circles are the actual points. Note the labels and the legend is added later (so running the code does not produce these).

2 Hoeffding's Inequality

Theorem 1.2 states:

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

We can rewrite the right side, as we know the a's and b's, to just n and we can divide by $1/n$ on both sides in the probability expression, so:

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

The expected value of the sum divided by n is the mean, so

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq \frac{\varepsilon}{n}\right\} \leq e^{-2\varepsilon^2/n}$$

That our right hand side in the probability expression is $\frac{\varepsilon}{n}$ it implies that the ε on the right side of the entire inequality is εn, so

$$\begin{aligned}\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq \varepsilon\right\} &\leq e^{-2(\varepsilon n)^2/n} \\ &= e^{-2\varepsilon n \varepsilon n/n} \\ &= e^{-2n\varepsilon^2}\end{aligned}$$

Which is what corollary 1.4 says. So we have proven we can go from equation 1.1 to equation 1.4 in the lecture notes.
This also holds for the second Hoeffding's inequality as it is just the probability expression multiplied by -1 on both sides (equation 1.2 to 1.5 in the lecture notes).

3 Illustration of Overfitting

3.1

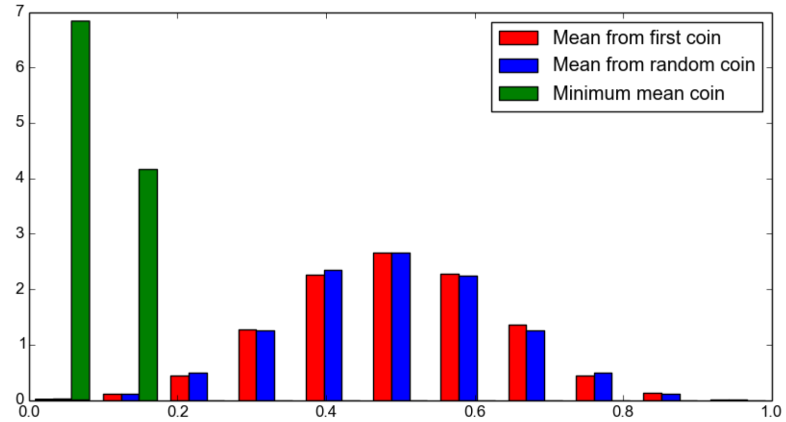
We can calculate $\mathbb{E}[\hat{\mu_i}]$, since we know the probabilities it takes 0 and 1 (1/2), as:

$$\begin{aligned}\mu_i &= \hat{\mu_i} \\ &= \frac{1}{10}\sum_{i=1}^{10} 1/2 \cdot 0 + 1/2 \cdot 1 \\ &= 1/2\end{aligned}$$

So the mean is 0.5.

3.2

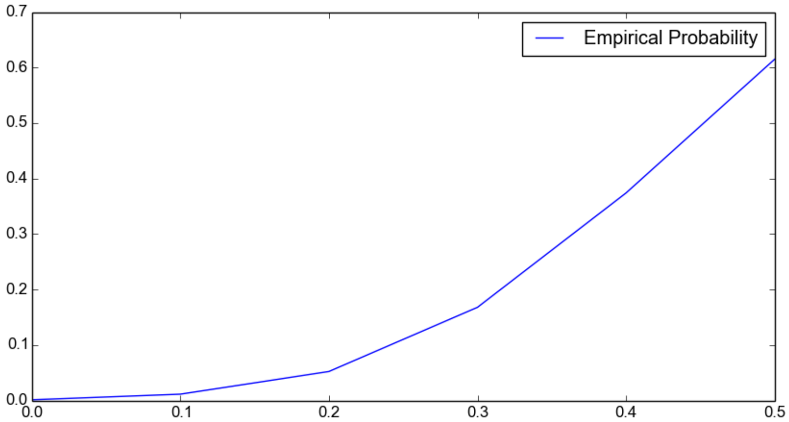
Running the code provides the following histogram:



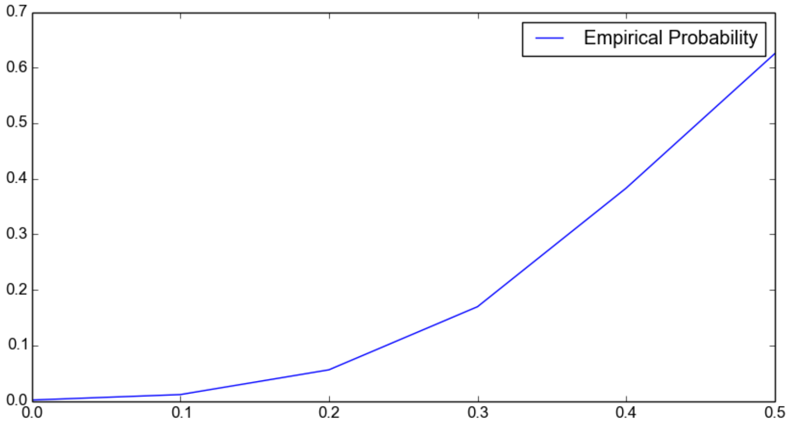
Where we can see the minimum coin only takes values in 0.0 and 0.1 and the other coins are spread around 0.5 as we would expect.

3.3

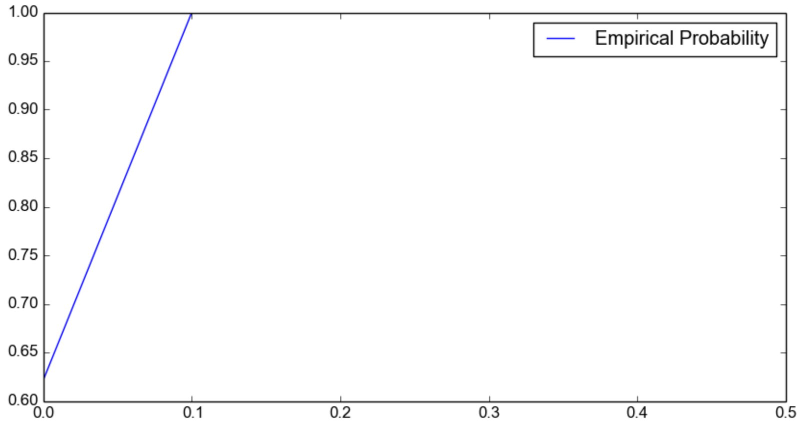
The empirical probability of $\hat{\mu}_1$ as function for x, where $x \in [0, 0.5]$ and x has a step size of 0.1 provides the following plot:



Doing the same for the random coins provides this plot:



Which looks almost exactly the same (as we would expect).
Doing the same again for the coin with minimum mean gives us:



which is 1 after 0.1 as there was not experiment where the minimum mean was larger than that.

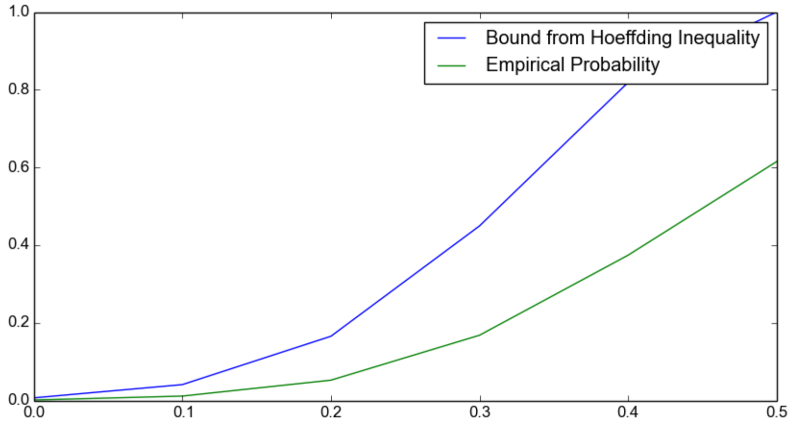
3.4

To calculate the bounds, we use corollary 1.4 from the lecture notes. So we have:

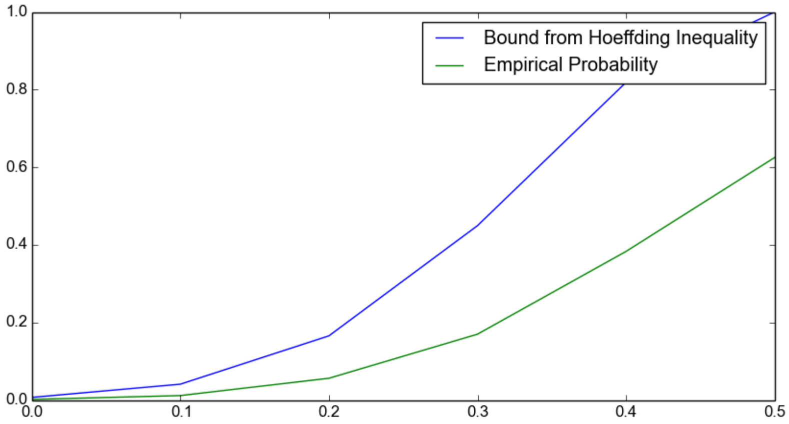
$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq \varepsilon\right\} \leq e^{-2n\varepsilon^2}$$

We actually need to find it in the cases where the left hand side in the probability expression is $[-0.5, 0]$ (and so is the epsilon). But we can actually just work in the positive interval $[0, 0.5]$.
Essentially, given an array `X=[0.0, 0.1, 0.2, 0.3, 0.4, 0.5]`, we know $n = 10$, so we compute the bound from Hoeffding's inequality as $e^{-2\cdot 10(0.5-X[i])^2}$ for $i = \{0, 1, .., 5\}$.

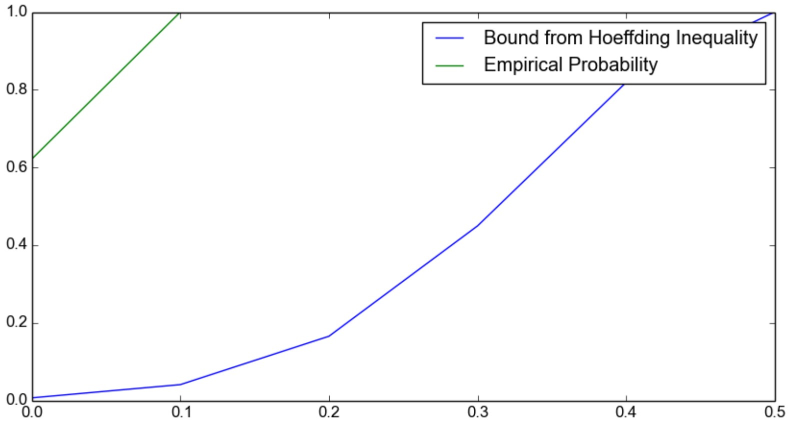
The following plot shows the comparison between the bound found using Hoeffding's inequality and the empirical probability for the first coin:



Likewise, we get the comparison between the random coin and the bound from Hoeffding's inequality:



And last it is compared to the coin with minimum mean:



Running the file `overfitting` will provide all the plots from question 3.

3.5

The previous point shows that the means $\hat{\mu_1}$ and $\hat{\mu_r}$ obeys Hoeffding's inequality as the probability is below the bound. The mean $\hat{\mu_m}$ does not as the empirical probability lies over the bound from Hoeffding's inequality.

4 Finite Hypothesis Set

4.1

The hypothesis space, $\mathcal{H}_1$, for the first approach is all combinations of $\mathcal{X} \times \{\text{male, female}\} \times \mathcal{Y}$.
So one example could be all males have a minor, but no females have a minor. Another could be all males have a minor except males that are 99 years old and females have no minor. So there are many hypotheses.

The hypothesis space, $\mathcal{H}_2$, for the second approach is all combinations of $(\{a_1, a_1+1, .., b_1\} \times \{\text{male}\} \times \{1\}) \times (\{a_2, a_2+1, .., b_2\} \times \{\text{female}\} \times \{1\})$, where $\{a_1, a_2, b_1, b_2\} \in \mathcal{X}$ and $a_1 \leq b_1$ and $a_2 \leq b_2$.
Since sets of 1 do not attribute to the total amount of hypotheses, we can actually write the hypothesis space $\mathcal{H}_2 : \{a_1, b_1\} \times \{a_2, b_2\}$ for $\{a_1, a_2, b_1, b_2\} \in \mathcal{X}$ and $a_1 \leq b_1$ and $a_2 \leq b_2$.
So one example could be all males in age range $[20, 30]$ have a minor, all female in age range $[20, 30]$ have a minor, and no one else has.

4.2

For the first approach it is $2^{101 \cdot 2} = 2^{202}$, since there are 202 "different" person (the combinations of age and gender), and the hypothesis assigns either 0 og 1 to these. So it all subsets where we assign 1 (or 0) to the persons. This is found by taking 2 to the power of the set size.

For the second approach there are $\frac{102*102}{2} \cdot \frac{101*102}{2} = 26532801$ hypotheses. The number of ranges we can make for ages 0 to 100 is equal to the triangle number to 101. Since we can have this range for both genders they are multiplied together. To compare, the number is between 2^{24} and 2^{25}.

4.3

Using equation 2.4 from the lecture notes, we can write the bound for $L(h)$ for the first approach as:

$$\begin{aligned} L(h) &\leq \hat{L}(h,S) + \sqrt{\frac{\ln \frac{M}{\delta}}{2n}} \\ &= \hat{L}(h,S) + \sqrt{\frac{\ln \frac{2^{202}}{\delta}}{2 \cdot 202}} \end{aligned}$$

where S is the sample set and δ is a certainty parameter as the inequality holds with probability $1 - \delta$. Likewise we can find the generalization bound for the second approach:

$$\begin{aligned} L(h) &\leq \hat{L}(h,S) + \sqrt{\frac{\ln \frac{M}{\delta}}{2n}} \\ &= \hat{L}(h,S) + \sqrt{\frac{\ln \frac{26532801}{\delta}}{2 \cdot 202}} \end{aligned}$$

so only the number of hypotheses, M, change.

4.4

The advantage of the first approach is it has more precise predictions when we have more hypotheses. The problem is the number of hypotheses. We need a lot of data to make the hypotheses representative. If we do not have enough data, the prediction rules become overly complex. So for example, our prediction could be 1 for a 14 year old (which in reality we know is very unlikely) because we only had one data point for that. Meanwhile we predict 0 for all $15, 16, 17, 18$ and 19 years old. This make for weird prediction rules.
The second approach is way more likely to be representative as we have a lot fewer hypotheses and is immediately preferable to use.

Another prediction problem could be problem of determining if a player is a rugby player with the input space of a weight (kg.) and an age, $\mathcal{X} = \{0, 1, .., 200\} \times \{0, 1, .., 100\}$. A preferable approach would be to use ranges for both things instead of single values just as before.