# 1 Illustration of Hoeffding's Inequality

This is some section!

# 2 The effect of scale (range) and normalization of random variables in Hoeffding's Inequality

Let all the assumptions of **corollary 2.4** be true for some random variables $X_1, ..., X_n$. Set $a_i = 0$ and $b_i = 1$ for all $i \in \{1, ..., n\}$. By the assumptions of **corollary 2.4** and our definition of $a_i$ and $b_i$, we now have that all the assumptions of **theorem 2.2** are true for $X_1, ..., X_n$. We can therefore use **theorem 2.2** to conclude that for all $\varepsilon > 0$

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

and

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

Since the assumptions of the corollary states that $\mathbb{E}(X_i) = \mu$ for all $i \in \{1, ..., n\}$, then we know that

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = n\mu \tag{3}$$

Since $b_i - a_i = 1$ for all $i \in \{1, ..., n\}$, we also know that

$$\sum_{i=1}^{n}(b_i - a_i)^2 = n \tag{4}$$

From line (XX-XX) we therefore get that for all $\varepsilon > 0$

$$\mathbb{P}\left\{\sum_{i=1}^{n} X_i - n\mu \geq \varepsilon\right\} \leq e^{-2\frac{\varepsilon^2}{n}} = e^{-2n\left(\frac{\varepsilon}{n}\right)^2} \tag{5}$$

and

$$\mathbb{P}\left\{\sum_{i=1}^{n} X_i - n\mu \leq -\varepsilon\right\} \leq e^{-2\frac{\varepsilon^2}{n}} = e^{-2n\left(\frac{\varepsilon}{n}\right)^2} \tag{6}$$

From this it clearly follows that for all $\varepsilon > 0$

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

and

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \leq -\frac{\varepsilon}{n}\right\} \leq= e^{-2n\left(\frac{\varepsilon}{n}\right)^2} \tag{8}$$

If $\varepsilon > 0$, then $\tilde{\varepsilon} = \frac{\varepsilon}{n} > 0$ for all $n \in \mathbb{N}$. We can therefore now conclude that for all $\tilde{\varepsilon} > 0$

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq \tilde{\varepsilon}\right\} \leq= e^{-2n\tilde{\varepsilon}^2} \tag{9}$$

and

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \leq -\tilde{\varepsilon}\right\} \leq= e^{-2n\tilde{\varepsilon}^2} \tag{10}$$

We have now proven that **corollary 2.4** follows from **theorem 2.2**.

# 3 Probability in Practice

This is some section!

# 4 Logistic Regression

This is some section!

## 4.1 Cross-entropy measure

Let $\mathcal{X}$ be some sample space, and let $\mathcal{Y}$ be the label space $\{-1, 1\}$, and assume that we want to learn the distribution of the labels $y$ conditioned on the value of a sample $x$, that is we want to learn the conditional probability $P(y|x)$ for $y \in -1, 1$ and all $x \in \mathcal{X}$. Also, assume that the distribution $P(y|x)$ can be parametrized by choosing $w$ in some parameter space $\mathcal{W}$. That is, by choosing $w \in \mathcal{W}$ we get the value of $P_w(y|x)$ for $y \in -1, 1$ and all $x \in \mathcal{X}$. In this context, the learning problem becomes to come up with a method for choosing some parameter $\hat{w} \in \mathcal{W}$ and hereby a corresponding distribution $P_{\hat{w}}(y|x)$, which somehow is our best guess of the true distribution of $y$ conditioned on $x$. The information we have available to base this choice on is some finite, labeled sample $S = \{(x_1, y_1), ..., (x_N, y_N)\}$, where each $y_i$ is assumed to have been sampled from $P_w(y|x_i)$ and all of them independently from each other.

The maximum likelihood method for choosing $\hat{w}$ solves this problem by defining the likelihood function $L_S$ for the given sample $S$ as

$$L_S(w) = \prod_{(x_n,y_n)\in\mathcal{S}} P_w(y_n|x_n) = \prod_{n=1}^{N} P_w(y_n|x_n) \tag{11}$$

and then saying that we should choose $\hat{w} \in \mathcal{W}$ such that $L_{\mathcal{S}}$ is maximized.

Since the function $-\ln$ is monotonically decreasing, this strategy is equivalent[1] to choosing $\hat{w} \in \mathcal{W}$ such that the function

$$f_S(w) = -\ln\left(\prod_{n=1}^{N} P_w(y_n|x_n)\right) = \sum_{n=1}^{N} (-\ln P_w(y_n|x_n)) \tag{12}$$

is minimized.

Since $y \in \{-1, 1\}$, then we can write $P_w(y|x)$ as

$$\begin{aligned} P_w(y|x) = & \quad (13)\\ [[y=1]]P_w(y=1|x) + [[y=1]]P_w(y=-1|x) = & \quad (14)\\ [[y=1]]h_w(x) + [[y=1]](1-h_w(x)) & \quad (15)\end{aligned}$$

where we simply have defined $h_w(x) = P(y = 1|x)$. We therefore have that

$$\begin{aligned} -\ln P_w(y_n|x_n) = & \quad (16)\\ -\ln([[y=1]]h_w(x) + [[y=-1]](1-h_w(x))) = & \quad (17)\\ [[y=1]](-\ln(h_w(x)) + [[y=-1]](-\ln(1-h_w(x))) = & \quad (18)\\ [[y=1]]\left(\ln\left(\frac{1}{h_w(x)}\right)\right) + [[y=-1]]\left(\ln\left(\frac{1}{1-h_w(x)}\right)\right) & \quad (19)\end{aligned}$$

[1] I define two strategies to be equivalent, if and only if they end up choosing the same $\hat{w}$ for all possible samples $S = \{(x_1, y_1), ..., (x_N, y_N)\}$.

By line (2) and line (6-9), we now get that

$$f_S(w) = \sum_{n=1}^{N} \left[ [[y_n = 1]] \left( \ln \left( \frac{1}{h_w(x_n)} \right) \right) + [[y_n = -1]] \left( \ln \left( \frac{1}{1 - h_w(x_n)} \right) \right) \right] \tag{20}$$

As I have already said, we will end up with the same $\hat{w}$ for a given sample $S$, if we minimize $f_S(w)$ as if we maximize $L_S(w)$. If we had started by saying that we would like to estimate the probability $h_w(x) = P_w(y = 1|x)$ by choosing $\hat{w}$ such that we minimize the error function $f_S(w)$ defined as in line (10), we would have ended up with the same estimates of $P(y|x)$ for $y \in \{-1, 1\}$ and $x \in \mathcal{X}$, as if we have used the maximum likelihood method. These two strategies are therefore equivalent.

## 4.2 Logistic regression loss gradient

In the algorithm for logistic regression, we determine the parameter $w \in \mathbb{R}^m$ in our final hypothesis

$$h_w(x) = P_w(y = 1|x) = \frac{e^{w^T x}}{1 + e^{w^T x}} \tag{21}$$

using a given labeled sample $S = \{(x_1, y_1), ..., (x_N, y_N)\}$ by minizing the function

$$E_{in}(w) = \frac{1}{N} \sum_{n=1}^{N} \ln(1 + e^{-y_n w^T x_n}) \tag{22}$$

We use gradient descent to do this, which means we have to find the loss gradient $\nabla E_{in}(w)$. We can use matrix calculus to to differentiate $E_{in}$

$$D_w(E_{in}(w)) = D_w \left( \frac{1}{N} \sum_{n=1}^{N} \ln(1 + e^{-y_n w^T x_n}) \right) = \tag{23}$$

$$\frac{1}{N} \sum_{n=1}^{N} D_w \left( \ln(1 + e^{-y_n w^T x_n}) \right). \tag{24}$$

Let us now focus on the inside of the sum

$$D_w \left( \ln(1 + e^{-y_n w^T x_n}) \right) = \frac{D_w(1 + e^{-y_n w^T x_n})}{1 + e^{-y_n w^T x_n}} = \tag{25}$$

$$\frac{D_w(e^{-y_n w^T x_n})}{1 + e^{-y_n w^T x_n}} = \frac{e^{-y_n w^T x_n} D_w(-y_n w^T x)}{1 + e^{-y_n w^T x_n}} = \tag{26}$$

$$-y_n x_n \frac{e^{-y_n w^T x_n}}{1 + e^{-y_n w^T x_n}} = -y_n x_n \Theta(-y_n w^T x_n) \tag{27}$$

Line (13-14) and line (15-17) now gives us that

$$D_w(E_{in}(w)) = \frac{1}{N}\sum_{n=1}^{N} D_w\left(\ln(1+e^{-y_n w^T x_n})\right) = \frac{1}{N}\sum_{n=1}^{N} -y_n x_n \Theta(-y_n w^T x_n) \quad (28)$$

which is what the exercise asked us to show.