# 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

This is some section!

# 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

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

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

$$f_S(w) = -\ln\left(\prod_{n=1}^{N} P_w(y_n|x_n)\right) = \sum_{n=1}^{N}(-\ln P_w(y_n|x_n)) \tag{2}$$

is minimized.

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

$$P_w(y|x) = \tag{3}$$
$$[[y=1]]P_w(y=1|x) + [[y=1]]P_w(y=-1|x) = \tag{4}$$
$$[[y=1]]h_w(x) + [[y=1]](1-h_w(x)) \tag{5}$$

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

$$-\ln P_w(y_n|x_n) = \tag{6}$$
$$-\ln([[y=1]]h_w(x) + [[y=-1]](1-h_w(x))) = \tag{7}$$
$$[[y=1]](-\ln(h_w(x)) + [[y=-1]](-\ln(1-h_w(x))) = \tag{8}$$
$$[[y=1]]\left(\ln\left(\frac{1}{h_w(x)}\right)\right) + [[y=-1]]\left(\ln\left(\frac{1}{1-h_w(x)}\right)\right) \tag{9}$$

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

$$f_S(w) = \sum_{n=1}^{N}\left[[[y_n=1]]\left(\ln\left(\frac{1}{h_w(x_n)}\right)\right) + [[y_n=-1]]\left(\ln\left(\frac{1}{1-h_w(x_n)}\right)\right)\right] \tag{10}$$

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.

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