Assignment n°1
Foundations of Machine Learning

Nicolas Bourriez

October 22, 2022

1 Probability and statistics

1.1 Bayes theorem

Let's define some random variables to describe our problem:

- $I = Infected$
- $S = Safe$
- $PT = PositiveTest$
- $NT = NegativeTest$

and the associated probabilities that we know:

- $P(I) = 1/1000$
- $P_S(PT) = 0.01$
- $P_S(NT) = 0.99$

We want to know the probability of a randomly tested person to be *actually* infected if its test is *positive* $P_{PT}(I)$.
We recall that the Bayes theorem gives us the following equality:

$$P_B(A) = \frac{P(A)P_A(B)}{P(B)} \tag{1}$$

so we can write

$$P_{PT}(I) = \frac{P(I)P_I(PT)}{P(PT)}$$

However we notice that

$$\begin{aligned} P(PT) &= P_I(PT)P(I) + P_S(PT)P(S) \\ &= 0.95 \times 0.001 + 0.01 \times 0.999 \\ &= \frac{1}{1000} \end{aligned}$$

So we write

$$\begin{aligned} P_{PT}(I) &= \frac{P(I)P_I(PT)}{0.011} \\ &= \frac{0.001 \times 0.95}{0.011} \\ &\simeq \frac{8.6}{1000} \end{aligned}$$

CONCLUSION: we find that the probability of a randomly tested person to be *actually* infected if its test is *positive* is equivalent to:

$$\boxed{\simeq \frac{8.6}{1000}}$$

1.2 Maximum Likelihood Estimator

1.2.1 Log-likelihood of the data

Given the N independent and identically distributed ($i.i.d$) samples $\mathbf{x} = x_1, x_2, ..., x_N$, we can write the *likelihood* of the data as

$$p(\mathbf{x}|\theta) = \mathcal{L}_\theta(\mathbf{x}) = \prod_{i=1}^{n} p(x_i|\theta) \qquad (i.i.d) \tag{2}$$

Using the natural logarithm ln, we thus write

$$\begin{aligned} l(\theta) = \ln(\mathcal{L}_\theta) &= \ln(\prod_{i=1}^{n} p(x_i|\theta)) \\ &= \sum_{i=1}^{n} \ln(p(x_i|\theta)) \\ &= \sum_{i=1}^{n} \ln(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x_i-\theta)^2}{\sigma^2}}) \qquad (w.r.t \quad \mu) \end{aligned}$$

We thus have the following equation using properties of ln

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{n} (-\ln\sqrt{2\pi\sigma^2} + \ln e^{-\frac{1}{2}\frac{(x_i-\theta)^2}{\sigma^2}}) \\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2 - \sum_{i=1}^{n}\frac{\ln(2\pi\sigma^2)}{2} \end{aligned}$$

CONCLUSION:

$$\boxed{l(\theta) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2 - \frac{n}{2}\ln(2\pi\sigma^2)} \tag{3}$$

1.2.2 Maximum Likelihood Estimator of μ

We define $\hat{\theta}(\mathbf{x})$ as the Maximum Likelihood Estimator of μ. It is thus the maximum argument of the likelihood $\mathcal{L}_\theta(\mathbf{x})$, which means that we want to take the likelihood for which the parameters θ provide the "highest" probabilities.

$$\hat{\theta}(\mathbf{x}) = \underset{\theta}{argmax} \quad \mathcal{L}_\theta(\mathbf{x})$$

Since we want to maximize our probability, we want to reach an extremum, and thus we can use the derivative of our *log-likelihood* by setting it to 0:

$$\hat{\theta} = \underset{\theta}{argmax} \quad l(\theta) \longrightarrow \frac{\partial l}{\partial \theta} = 0 \qquad (w.r.t \quad \mu)$$

We thus first compute the partial derivative of l w.r.t θ, which can be written as being proportional to an easier expression from .Equation (3)

$$\begin{aligned} \frac{\partial l(\theta)}{\partial \theta} &= \frac{\partial l}{\partial \theta}\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2 - \frac{n}{2}\ln(2\pi\sigma^2))\right] \\ \frac{\partial l(\theta)}{\partial \theta} &\propto \sum_i (\theta - x_i) \\ &\propto n\theta - \sum_i x_i \end{aligned}$$

Thus we can write:

$$\begin{aligned} \frac{\partial l}{\partial \theta} = 0 &\iff n\theta - \sum_i x_i = 0 \\ &\iff \theta = \frac{1}{n}\sum_i x_i \end{aligned}$$

CONCLUSION: We find that the Maximum Likelihood Estimator of μ is equal to the mean $\overline{X}_n$

$$\boxed{\hat{\theta} = \underset{\theta}{argmax} l(\theta) = -\frac{1}{n}\sum_i x_i} \tag{4}$$

1.2.3 Maximum Likelihood Estimator of σ^2

We define $\hat{\theta}(\mathbf{x})$ as the Maximum Likelihood Estimator of σ^2.
Similarly to 1.2.2, we can write the log-likelihood of θ with respect to σ^2:

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{n} \ln(\frac{1}{\sqrt{2\pi\theta}} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\theta}}) \qquad (w.r.t \quad \sigma^2) \\ &= -\frac{1}{2\theta}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{n}{2}\ln(2\pi\theta) \end{aligned}$$

Same as for the MLE of μ, here we want to find the right σ^2 that will maximize the likelihood $\mathcal{L}_\theta(\mathbf{x})$.
We thus have:

$$\hat{\theta} = \underset{\theta}{argmax} \quad l(\theta) \longrightarrow \frac{\partial l}{\partial \theta} = 0 \qquad (w.r.t \quad \sigma^2)$$

Computing the partial derivative of l, we get:

$$\begin{aligned} \frac{\partial l}{\partial \theta} &= \frac{\left[-\sum_{i=1}^{n}(x_i-\mu)^2\right]' 2\theta - \left[-\sum_{i=1}^{n}(x_i-\mu)^2\right](2\theta)'}{(2\theta)^2} - \left[(\frac{n}{2})'\ln(2\pi\theta) + \frac{n}{2}\ln(2\pi\theta)'\right] \\ &= \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\theta^2} - \frac{n}{2\theta} \\ &= \frac{\sum_{i=1}^{n}(x_i-\mu)^2 - n\theta}{2\theta^2} \qquad (\theta \neq 0) \end{aligned}$$

We pose

$$\hat{\theta} = \underset{\theta}{argmax} \quad l(\theta) \longrightarrow \frac{\partial l}{\partial \theta} = 0 \qquad (w.r.t \quad \sigma^2)$$

And thus we can then compute the MLE of σ^2 which gives:

$$\begin{aligned} \hat{\theta} = \underset{\theta}{argmax} l(\theta) &\longrightarrow \frac{\sum_{i=1}^{n}(x_i-\mu)^2 - n\theta}{2\theta} = 0 \\ &\longrightarrow \sum_{i=1}^{n}(x_i-\mu)^2 - n\theta = 0 \\ &\longrightarrow \theta = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 \end{aligned}$$

CONCLUSION: We find that the Maximum Likelihood Estimator of σ^2 is equal to :

$$\boxed{\hat{\theta} = \underset{\theta}{argmax} l(\theta) = \frac{1}{n}\sum_i (x_i-\mu)^2} \tag{5}$$

1.2.4 Bonus question

We recall that an unbiased estimator $\hat{\theta}$ is equal to $\mathbb{E}(\hat{\theta}) = \theta$.
We want to show that the Maximum Likelihood Estimator of σ^2 is **biased**.
From .Equation (5), we already have the following expression:

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

Moreover since we recall from .Equation (4) that $\hat{\theta}$ with respect to μ is the ***mean*** $\overline{X}_n$, we can write:

$$\begin{aligned}
\hat{\theta} &= \frac{1}{n}\sum_{i=1}^{n}(X_i - 2X_i\overline{X}_n + \overline{X}_n^2) \\
&= \frac{1}{n}(\sum_{i=1}^{n} X_i^2 - 2\overline{X}_n\sum_{i=1}^{n} X_i + \sum_{i=1}^{n}\overline{X}_n^2) \\
&= \frac{1}{n}(\sum_{i=1}^{n} X_i^2 - 2n\overline{X}_n^2 + n\overline{X}_n^2) \qquad (n\overline{X_n} = \sum_{i=1}^{n} X_i)
\end{aligned}$$

Using the expression we just found, and using the linearity of the Expected Value, we get:

$$\mathbb{E}(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i^2) - \mathbb{E}(\overline{X}_n^2)$$

Let's look further into each element of $\mathbb{E}(\hat{\theta})$:

- $\mathbb{E}(X_i^2)$?

  We know that

  $$\begin{aligned}
  \sigma^2 = \mathbb{V}(X_i) &= \mathbb{E}((X_i - \mathbb{E}(X_i)^2) \\
  &= \mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2
  \end{aligned}$$

  Thus we end having:

  $$\mathbb{E}(X_i^2) = \sigma^2 + \mu^2$$

  And similarly

  $$\mathbb{E}(\overline{X}_n) = \mu$$

- $\mathbb{E}(\overline{X}_n^2)$?

  We know that:

  $$\begin{aligned}
  \mathbb{V}(\overline{X}_n) &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(X_i) \\
  &= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 \\
  &= \frac{1}{n^2}\sigma^2 \\
  &= \frac{\sigma^2}{n}
  \end{aligned}$$

  We can now compute $\mathbb{E}(\overline{X}_n^2)$:

  $$\begin{aligned}
  \mathbb{E}(\overline{X}_n^2) &= \mathbb{V}(\overline{X}_n) + (\mathbb{E}(\overline{X}_n))^2 \\
  &= \frac{\sigma^2}{n} + \mu^2
  \end{aligned}$$

Aggregating the two expressions we got, we can now formulate $\mathbb{E}(\hat{\theta})$:

$$\begin{aligned}
\mathbb{E}(\hat{\theta}) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i^2) - \mathbb{E}(\overline{X}_n^2) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\sigma^2+\mu^2) - \frac{\sigma^2}{n} + \mu^2 \\
&= \frac{1}{n}\times n(\sigma^2+\mu^2) - \frac{\sigma^2}{n} - \mu^2 \\
&= \sigma^2 + \mu^2 + \frac{\sigma^2}{n} - \mu^2 \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}$$

CONCLUSION:$\mathbb{E}(\hat{\theta})$ is a ***biased*** estimator of σ^2 since $\mathbb{E}(\hat{\theta}) \neq \sigma^2$, and its bias is equal to :

$$\begin{aligned}
b &= \mathbb{E}(\hat{\theta}) - \theta \\
&= \frac{n-1}{n}\sigma^2 - \sigma^2 \\
&\boxed{b = -\frac{\sigma^2}{n}}
\end{aligned} \tag{6}$$

2 Linear Regression

2.1 Parameters of a linear regression

2.1.1 Derivative of w

We have $\mathbf{w} = (w_1, w_2)^T = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$ so $\hat{y}_n = w_1 x_{1,n} + w_2 x_{2,n}$

We know from the exercise details that our loss function that we want to derive is the following:

$$J(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2$$

We now compute the partial derivatives of J:

- $\frac{\partial J(\mathbf{w})}{\partial w_1}$

$$\frac{\partial J(\mathbf{w})}{\partial w_1} = \frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial w_1}(y_n - \hat{y}_n)^2$$

Focusing only on the term $\frac{\partial}{\partial w_1}(y_n - \hat{y}_n)^2$:

$$\begin{aligned}
\frac{\partial}{\partial w_1}(y_n - \hat{y}_n)^2 &= (2y_n - \hat{y}_n)\frac{\partial}{\partial w_1}(y_n - \hat{y}_n) && (chain \quad rule) \\
&= (2y_n - \hat{y}_n) \times (-x_{1,n}) && (-x_{1,n} = \frac{\partial}{\partial w_1}(y_n - w_1 x_{1,n} - w_2 x_{2,n})
\end{aligned}$$

We end up with the following expression:

$$\begin{aligned}
\frac{\partial J(\mathbf{w})}{\partial w_1} &= \frac{2}{N}\sum_{n=1}^{N}(y_n - w_1 x_{1,n} - w_2 x_{2,n}) \times -x_{1,n} \\
&= \frac{2}{N}(w_1\sum_{n=1}^{N}x_{1,n}^2 + w_2\sum_{n=1}^{N}x_{1,n}x_{2,n} - \sum_{n=1}^{N}x_{1,n}y_n)
\end{aligned}$$

- $\frac{\partial J(\mathbf{w})}{\partial w_2}$ Similarly, we get:

$$\begin{aligned}
\frac{\partial J(\mathbf{w})}{\partial w_1} &= \frac{2}{N}\sum_{n=1}^{N}(y_n - w_1 x_{1,n} - w_2 x_{2,n}) \times -x_{2,n} \\
&= \frac{2}{N}(w_1\sum_{n=1}^{N}x_{2,n}^2 + w_2\sum_{n=1}^{N}x_{1,n}x_{2,n} - \sum_{n=1}^{N}x_{2,n}y_n)
\end{aligned}$$

To simplify the notation, we can write the following terms as so:

- $t_1 = \sum_{n=1}^{N} x_{1,n}^2$
- $t_2 = \sum_{n=1}^{N} x_{2,n}^2$
- $s = \sum_{n=1}^{N} x_{1,n}x_{2,n}$
- $v_1 = \sum_{n=1}^{N} x_{1,n}y_n$
- $v_2 = \sum_{n=1}^{N} x_{2,n}y_n$

And get

$$\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = 0 \iff \begin{cases} \frac{\partial J(\mathbf{w})}{\partial w_1} = 0 \\ \frac{\partial J(\mathbf{w})}{\partial w_2} = 0 \end{cases}$$
$$\iff \begin{cases} w_1 = \frac{t_2 v_1 - s v_2}{t_1 t_2 - s^2} \\ w_2 = \frac{t_1 v_2 - s v_2}{t_1 t_2 - s^2} \end{cases}$$

CONCLUSION: The derivative of $\mathbf{w}$ when set to 0 is equal to :

$$\boxed{\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = 0 \iff \begin{cases} w_1 = \frac{t_2 v_1 - s v_2}{t_1 t_2 - s^2} \\ w_2 = \frac{t_1 v_2 - s v_2}{t_1 t_2 - s^2} \end{cases}} \tag{7}$$

2.1.2 Compatibility with closed-form solution

Since we are in dimension 2, we can see that $\mathbf{X} = (\sum\limits_{n=1}^{N} x_{1,n} \sum\limits_{n=1}^{N} x_{2,n})$ and thus we get:

$$\mathbf{X^T X} = \begin{pmatrix} \sum\limits_{n=1}^{N} x_{1,n}^2 & \sum\limits_{n=1}^{N} x_{1,n}x_{1,n} \\ \sum\limits_{n=1}^{N} x_{1,n}x_{1,n} & \sum\limits_{n=1}^{N} x_{2,n}^2 \end{pmatrix}$$

Using the formula to compute the inverse of a matrix, we can find:

$$\mathbf{X^T X}^{-1} = \frac{1}{t_1 t_2 - s^2} \begin{pmatrix} t_2 & -s \\ -s & t_1 \end{pmatrix}$$

CONCLUSION: Multiplying by $\mathbf{X^T y}$, we end up with the following equation:

$$\boxed{\mathbf{w} = (\mathbf{X^T X})^{-1}\mathbf{X^T y} = \frac{1}{t_1 t_2 - s^2} \begin{pmatrix} t_2 v_1 - s v_2 \\ t_1 v_2 - c v_1 \end{pmatrix}} \tag{8}$$

which appear coherent

2.2 Least Square Loss

2.2.1 Log-likelihood of the observations

- **Likelihood**
  We know that $D = (x_1, y_1), ..., (x_n, y_n)$, thus we can write the likelihood of D given a set of parameters $\theta = (w, \sigma^2)$ as

  $$p(D|\theta) = \prod_{n=1}^{N} p(x_n, y_n | w, \sigma^2) \qquad (i.i.d)$$

  We recall that $p(a, b) = p(a|b)p(b)$, hence

  $$p(D|\theta) = \prod_{n=1}^{N} p(x_n | y_n, w, \sigma^2) p(y_n, w, \sigma^2)$$

But we also recall according to the compound probability theorem that

$$p(A \cap B) = p(A|B)p(B) = p(B|A)p(A)$$

We can thus write the likelihood as so:

$$p(D|\theta) = \prod_{n=1}^{N} p(y_n|x_n, w, \sigma^2)p(x_n, w, \sigma^2)$$

- **Log-likelihood**
  We wish to write the log-likelihood of our observations D as a function of w.
  Using the natural log:

$$l(w) = \ln\left[\prod_{n=1}^{N} p(y_n|x_n, w, \sigma^2)p(x_n, w, \sigma^2)\right]$$
$$= \sum_{i=1}^{N}\left[\ln(p(y_n|x_n, w, \sigma^2)) + \ln(p(x_n|w, \sigma^2))\right]$$

  Analysing each term of the sum:

  – $\sum_{i=1}^{N} \ln(p(x_n|w, \sigma^2)) = \sum_{i=1}^{N} \ln[\frac{1}{\sqrt{2\pi\sigma^2}} e^{(-\frac{(x_n-w)^2}{2\sigma^2})}] = -N\ln(\sqrt{2\pi\sigma^2}) - \frac{1}{\sigma^2}\frac{1}{2}\sum_{n=1}^{N}(x_n - w)^2$

  – $\sum_{i=1}^{N} \ln(p(y_n|x_n, w, \sigma^2)) = \sum_{i=1}^{N} \ln[\frac{1}{\sqrt{2\pi\sigma^2}} e^{(-\frac{(y_n-w^T x_n)^2}{2\sigma^2})}] = -N\ln(\sqrt{2\pi\sigma^2}) - \frac{1}{\sigma^2}\frac{1}{2}\sum_{n=1}^{N}(y_n - w^T x_n)^2$

CONCLUSION:

$$\boxed{l(w) = -2N\ln(\sqrt{(}2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - w)^2 - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - w^T x_n)^2} \tag{9}$$

2.2.2 Maximization of log-likelihood

We notice first that due to **monotony** of ln, we have:

$$\underset{w}{argmax} \quad p(D|w, \sigma^2) \Longrightarrow \underset{w}{argmax} \ln(p(D|w, \sigma^2))$$

Since we want to maximize our log-likelihood with respect to w, we can set:

$$\hat{w} = \underset{w}{argmax} \quad l(w) \Longrightarrow \frac{\partial l(w)}{\partial w} = 0$$

Computing the partial derivative of $l(w)$:

$$\frac{\partial l(w)}{\partial w} = \frac{\partial}{\partial w}\left[-2N\ln(\sqrt{(}2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - w)^2 - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - w^T x_n)^2\right]$$

- $-2N\ln(\sqrt{(}2\pi\sigma^2)$ is a constant, so its derivative is equal to 0
- $\frac{\partial}{\partial w}\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - w)^2\right] \propto Nw - \sum_n x_n$ (from Exercise 1.2.2)

We arrive then to the following expression:

$$\frac{\partial l(w)}{\partial w} = 0 \Longleftrightarrow Nw - \sum_n x_n - \frac{\partial}{\partial w}\left[\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - w^T x_n)^2\right]$$
$$\Longleftrightarrow N\left[w - \frac{1}{N}\sum_n x_n - \frac{\partial}{\partial w}\left[\frac{1}{2\sigma^2}\frac{1}{N}\sum_n(y_n - w^T x_n)^2\right]\right]$$

CONCLUSION: We recognize that to *maximize* our log-likelihood, we have to *minimize* the following term, which is the sum of the squared errors:

$$\boxed{\frac{1}{N}\sum_n (y_n - w^T x_n)^2} \tag{10}$$

3 Logistic Regression

3.1 Link between odd ratio and sigmoid function

As per its definition, a *sigmoid function* is represented as follows

$$\sigma : x \mapsto \frac{1}{1+e^{-x}}$$

We want to prove that using a linear model $\mathbf{w^T x}$,

$$\sigma(\mathbf{w^T x}) = \frac{1}{1+e^{-\mathbf{w^T x}}} \iff \mathbf{w^T x} \sim \log(\frac{p(y=1)}{p(y=0)})$$

Starting from the right-hand side of the equivalence and applying the exp on the natural logarithm, we have:

$$\begin{aligned}
&\iff e^{\mathbf{w^T x}} \sim \frac{p(y=1)}{p(y=0)} \\
&\iff \frac{1}{e^{\mathbf{w^T x}}} \sim \frac{p(y=0)}{p(y=1)} \\
&\iff 1 + \frac{1}{e^{\mathbf{w^T x}}} \sim 1 + \frac{p(y=0)}{p(y=1)} \\
&\iff 1 + e^{(-\mathbf{w^T x})} \sim \frac{p(y=1)+p(y=0)}{p(y=1)} \\
&\iff \frac{1}{1+e^{-\mathbf{w^T x}}} \sim \frac{p(y=1)}{p(y=1)+p(y=0)}
\end{aligned}$$

CONCLUSION: We observe that predicting the log of the odd ratio with a linear model is equivalent to predicting the proba of *winning* by applying a sigmoid function to the output of a linear model, which translates to:

$$\boxed{\frac{1}{1+e^{\mathbf{w^T x}}} \sim \frac{p(y=1)}{p(y=1)+p(y=0)}} \tag{11}$$

3.2 Link between maximization of log-likelihood and logistic regression model

Given a set of observations $D = \{(x_1, y_1), ..., (x_n, y_n)\}$, we can write the likelihood of this set as

$$p(D) = \prod_{n=1}^{N} p(x_n, y_n) \qquad (i.i.d)$$

We recall that $p(a, b) = p(a|b)p(b) = p(b|a)p(a)$, hence

$$p(D) = \prod_{n=1}^{N} p(y_n|x_n)p(x_n)$$

Observing the odd ratio, we also recognize that the probability mass function of the random variable X is the one of a Bernoulli variable:

$$P(X=x) = \begin{cases} p \quad if \quad x=1 \\ 1-p \quad if \quad x=0 \\ 0 \quad otherwise \end{cases}$$

which also translate into the following expression:

$$\begin{aligned}
P(X=x) &= p^x(1-p)^{(1-x)}, \quad x \in 0,1 \\
&= e^{x \ln p} e^{(1-p)\ln(1-x)} \qquad (a^n = e^{n \ln(a)})
\end{aligned}$$

From .Equation (11), we have

$$P(Y=1) = \frac{1}{1+e^{-\mathbf{w^T x}}}$$

Using the probability distribution we observed previously, we get

$$P(Y=y)=\begin{cases}\frac{1}{1+e^{-\mathbf{w^Tx}}} & if \quad y=1\\ 1-\frac{1}{1+e^{-\mathbf{w^Tx}}} & if \quad y=0\\ 0 \quad otherwise\end{cases}$$

Partial conclusion: the PMF of y ***depends*** on $\mathbf{w}$.

We now apply the natural logarithm to get the *log-likelihood* of the set of observations D:

$$\ln(p(D))=\ln\prod_{n=1}^{N}p(y_n|x_n)p(x_n)$$
$$\ln(p(D))=\sum_{n=1}^{N}\ln(p(y_n|x_n))+\sum_{n=1}^{N}\ln(p(x_n))$$

If we decompose each term of our expression, we then have

- For $\sum_{n=1}^{N}\ln(p(y_n|x_n))$

$$\begin{aligned}\sum_{n=1}^{N}\ln(p(y_n|x_n))&=\sum_{n=1}^{N}\ln(e^{y_n\ln x_n}e^{(1-x_n)\ln(1-y_n)})\\&=\sum_{n=1}^{N}[y_n\ln x_n+(1-x_n)\ln(1-y_n)]\end{aligned}$$

- For $\sum_{n=1}^{N}\ln(p(x_n))$, similarly we get

$$\begin{aligned}\sum_{n=1}^{N}\ln(p(x_n))&=\sum_{n=1}^{N}\ln(p(x_n|p))\\&=\sum_{n=1}^{N}\ln(e^{x_n\ln p}e^{(1-x_n)\ln(1-p)})\\&=\sum_{n=1}^{N}[x_n\ln p+(1-p)\ln(1-x_n)]\end{aligned}$$

We can recall that maximizing the *log-likelihood* of $p(D)$ gives us the following expression:

$$\underset{w}{argmax} \quad \ln(p((D))\Longrightarrow\frac{\partial\ln(p((D))}{\partial w}=0$$

with the partial derivative of the *log-likelihood* being

$$\begin{aligned}\frac{\partial\ln(p((D))}{\partial w}=\frac{\partial}{\partial w}&\left[\sum_{n=1}^{N}y_n\ln x_n(1-x_n)\ln(1-y_n)\right]\\&+\\&\frac{\partial}{\partial w}\left[\sum_{n=1}^{N}x_n\ln p+(1-p)\ln(1-x_n)\right]\quad(\lambda f+\mu g)'=\lambda f'+\mu g')\end{aligned}$$

But we remember that only the PMF of y depends on $\mathbf{w}$, so we deduct that the partial derivative $\frac{\partial}{\partial\mathbf{w}}$ of $\sum_{n=1}^{N}\ln(p(x_n))$ is a ***constant***.

CONCLUSION: We then recognize that we have a logistic regression model and its corresponding *training loss* by minimizing the following expression:

$$\boxed{J(\mathbf{w})=-\sum_{n=1}^{N}[y_n\ln x_n(1-x_n)\ln(1-y_n)]=-\sum_{n=1}^{N}\ln(p(y_n|x_n))} \qquad (12)$$

4 Clustering

4.1 K-Means

$$If \quad a_{n,k} \in k, then \quad a_{n,k} = 1 \tag{13}$$
$$else \quad a_{n,k} = 0 \tag{14}$$

Since we want to minimize the inertia, we want that:

$$\frac{\partial J(c_k)}{\partial c_k} = 0$$

Writing the partial derivative of $J(c_k)$:

$$\begin{aligned}
\frac{\partial J(c_k)}{\partial c_k} &= \frac{1}{N}\frac{\partial}{\partial c_k}\left[\sum_{n=1}^{N}\sum_{k=1}^{N} a_{n,k} \parallel \mathbf{x}_n - \mathbf{c}_k \parallel_2^2\right] \\
&= \frac{1}{N}\frac{\partial}{\partial c_k}\left[\sum_{n=1}^{N}\sum_{k=1}^{N} a_{n,k}(\mathbf{x}_n - \mathbf{c}_k)^T(\mathbf{x}_n - \mathbf{c}_k)\right] \\
&= \frac{1}{N}\frac{\partial}{\partial c_k}\left[\sum_{n=1}^{N}\sum_{k=1}^{N} a_{n,k}(\mathbf{x}_n^T\mathbf{x}_n - \mathbf{x}_n^T\mathbf{c}_k - \mathbf{c}_k^T\mathbf{x}_n + \mathbf{c}_k^T\mathbf{c}_k)\right] \\
&= \frac{1}{N}\sum_{n=1}^{N} a_{n,k}\frac{\partial}{\partial c_k}(\mathbf{x}_n^T\mathbf{x}_n - \mathbf{x}_n^T\mathbf{c}_k - \mathbf{c}_k^T\mathbf{x}_n + \mathbf{c}_k^T\mathbf{c}_k) \qquad (Eq.(14) : a_{n,k} \notin k \rightarrow a_{n,k} = 0)
\end{aligned}$$

Because $(uv)' = (u'v + uv')$ and $\frac{\partial(x^T a)}{\partial a} = \frac{\partial(a^T x)}{\partial a} = a$, we thus get:

$$\frac{\partial J(c_k)}{\partial c_k} = \frac{1}{N}\sum_{n=1}^{N} a_{n,k}(-2\mathbf{x}_n + 2\mathbf{c}_k)$$

We can now compute:

$$\begin{aligned}
\frac{\partial J(c_k)}{\partial c_k} = 0 &\Longleftrightarrow \frac{1}{N}\sum_{n=1}^{N} a_{n,k}(-2\mathbf{x}_n + 2\mathbf{c}_k) = 0 \\
&\Longleftrightarrow \frac{2}{N}\sum_{n=1}^{N}(-a_{n,k}\mathbf{x}_n + a_{n,k}\mathbf{c}_k) = 0 \\
&\Longleftrightarrow \frac{2}{N}\left[\sum_{n=1}^{N} a_{n,k}\mathbf{c}_k - \sum_{n=1}^{N} a_{n,k}\mathbf{x}_n\right] = 0 \\
&\Longleftrightarrow \mathbf{c}_k = \frac{\sum\limits_{n=1}^{N} a_{n,k}\mathbf{x}_n}{\sum\limits_{n=1}^{N} a_{n,k}}
\end{aligned}$$

CONCLUSION: When minimizing J, cluster centers $c_1, ..., c_K$ are the means of the points assigned to the respective clusters as described by:

$$\boxed{\frac{\partial J(c_k)}{\partial c_k} = 0 \Longleftrightarrow \mathbf{c}_k = \frac{\sum\limits_{n=1}^{N} a_{n,k}\mathbf{x}_n}{\sum\limits_{n=1}^{N} a_{n,k}}} \tag{15}$$

4.2 Hierarchical clustering and Levensthein distance

We want to show that the Levensthein distance is indeed a *distance*, using the mathematical definition of the term.

Reasoning by the absurd
If the Levenshtein distance was ***not*** a general distance, thus according to the mathematical definition:

$$\begin{cases} \forall (a,b) \in E^2, d(a,b) \neq d(b,a) \\ \forall (a,b) \in E^2, d(a,b) = 0 \not\equiv a = b \\ \forall (a,b,c), (d(a,c) > d(a,b) + d(b,c) \end{cases}$$

*Is the Levenshtein distance respecting these criterion, and thus **is not** a general distance ?*

- **Symmetry**
Given two strings of characters $a =$ "*machine*" and $b =$ "*learning*" with $(a,b) \in E^2, E = a, b, ..., z$ and d the Levenshtein distance:

$$\begin{cases} d(a,b) = d(\text{"}machine\text{"}, \text{"}learning\text{"}) = 5 \qquad (\text{"}m\text{"}, \text{"}c\text{"}, \text{"}h\text{"}, \text{"}r\text{"}, \text{"}g\text{"}) \\ d(b,a) = d(\text{"}learning\text{"}, \text{"}machine\text{"}) = 5 \qquad (\text{"}m\text{"}, \text{"}c\text{"}, \text{"}h\text{"}, \text{"}r\text{"}, \text{"}g\text{"}) \end{cases}$$

$$\boxed{d(a,b) = d(b,a)}$$

- **Separation** Given two strings $a =$ "*hello*" and $b =$ "*hello*", with $(a,b) \in E^2, E = a, b, ..., z$ and d the Levenshtein distance:

$$\boxed{d(a,b) = d(b,a) = 0}$$

- **Triangular inequality** Given three strings $a =$ "*hello*", $b =$ "*hallo*" and $c =$ "*hola*, with $(a,b,c) \in E^3, E = a, b, ..., z$ and d the Levenshtein distance:

$$\begin{cases} d(a,c) = d(\text{"}hello\text{"}, \text{"}hola\text{"}) = 3 \qquad (\text{"}e\text{"}, \text{"}l\text{"}, \text{"}a\text{"}) \\ d(a,b) + d(b,c) = 1 + 3 = 4 \end{cases}$$

$$\boxed{d(a,c) \not> d(a,b) + d(b,c)}$$

CONCLUSION: **Reasoning by the *absurd***, we clearly observe that all of the three criterias for NOT being a general distance are countered by the Levenshtein distance, hence we can say that the Levenshtein distance is a ***general distance***.

$$\boxed{Levenshtein \quad distance = \begin{cases} \forall (a,b) \in E^2, d(a,b)d(b,a) \\ \forall (a,b) \in E^2, d(a,b) = 0 \equiv a = b \\ \forall (a,b,c), (d(a,c) \leq d(a,b) + d(b,c) \end{cases}} \tag{16}$$

4.3 Single-linkage criterion

Given two clusters C_1 and C_2, with $C_1 = (x,y), C_2 = (x,z) \quad \forall (x,y) \in E^2, E \subset \mathbb{R}$ and $\mathbf{D}(C_1, C_2) = \min_{\mathbf{x}_1 \in C_1, \mathbf{x}_2 \in C_2,} d(x_1, x_2)$ being the ***single-linkage*** criterion

- *Is the Symmetry criterion respected ?*
$\mathbf{D}(C_1, C_2) = \min_{\mathbf{x}_1 \in C_1, \mathbf{x}_2 \in C_2,} d(x_1, x_2) = 0$ but $x_1 \neq x_2$

CONCLUSION: Using a counter example, we proved that the single-linkage criterion **isn't** a *general* distance.