Introduzione alla data science e al pensiero computazionale
Lezione 11: Regressione Logistica

Giancarlo Succi
Dipartimento di Informatica – Scienza e Ingegneria
Università di Bologna
g.succi@unibo.it

Content

- Covariance
- Correlation (aka Pearson product-moment correlation coefficient)
- Relationship between Pearson correlation and linear regression

Outline

- Likelihood function, definition
- Maximum likelihood
- Log likelihood
- Logistic regression

Some slides are take from:
https://www.cs.ox.ac.uk/people/nando.defreitas/

Likelihood function

Let $X_1, X_2, ..., X_n$ denote a random sample from p.d.f.

$$X_i \sim f_\theta(x),$$

where θ represents one ore more unknown parameters of the distribution.
The joint p.d.f. of $X_1, X_2, ..., X_n$ is $f_\theta(x_1), f_\theta(x_2), ..., f_\theta(x_n)$.
If we consider this joint p.d.f. as a function of θ it is called *likelihood function* of a random sample:

$$L_{x_1,x_2,...,x_n}(\theta) = f_\theta(x_1), f_\theta(x_2), ..., f_\theta(x_n).$$

Maximum likelihood

Let's consider an estimator of θ:

$$\hat{\theta} = u(X_1, X_2, ..., X_n).$$

If for every possible θ $L_{x_1,x_2,...,x_n}(\hat{\theta})$ is at least as great as $L_{x_1,x_2,...,x_n}(\theta)$ then $\hat{\theta}$ is called *maximum likelihood estimator*.
Finally:

$$\hat{\theta} =_\theta (L_{x_1,x_2,...,x_n}(\theta))$$

Maximum loglikelihood

Note that, since the likelihood function $L_{x_1,x_2,...,x_n}(\theta)$ and its logarithm $ln(L_{x_1,x_2,...,x_n}(\theta))$, are maximized for the same value θ, either likelihood or its logarithm can be used to find maximum likelihood estimator:

$$\hat{\theta} =_{\theta} (ln(L_{x_1,x_2,...,x_n}(\theta)))$$

The concept of regression

Regressions can be of multiple types, so far we have analysed the so called OLS regression:

- quadratic cost function of the kind $\sum_i (\hat{y}_i - y_i)^2$
- linear model of the kind $\hat{y} = \mathbf{A}\mathbf{x} + \eta$

What if:

- we use a different objective function, or
- we use a different model

?

Remember that model is called "the mean function" and its inverse "the link function."

Posing a different problem

Let's suppose to have:

- three iid random variables y_i with $i \in [1 \dots 3]$
- with the same partially unknown pdf, that is
- $(\forall i)\ y_i \sim N(\theta, 1)$
- θ to be determined.

We want to determine the value of θ that maximizes the probability of obtaining y_1 and y_2 and y_3.

In other terms our objective function is the probability of occurrence of y_1 and y_2 and y_3.

We are looking for a maximum likelihood estimator!

Computing the highest probability

Our objective function is therefore:

$$P(y_1, y_2, y_3|\theta) = P(y_1|\theta) \times P(y_2|\theta) \times P(y_3|\theta)$$

We can rewrite this problem as:

$$\max_{\theta}(\prod_{i=1}^{3} P(y_i|\theta))$$

Note that since θ is a *crisp* value:

$$y_i \sim N(\theta, 1) = \text{a shift of } \theta \text{ of } N(0, 1)$$

Using concrete numbers - 1

Let us assume that:

- $y_1 = 1$
- $y_2 = 0.5$
- $y_3 = 1.5$

Remember that $N(\theta, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\theta)^2}{2\sigma^2}}$

Therefore, we want to maximize:

$$\prod_{i=1}^{3} P(y_i|\theta) = \prod_{i=1}^{3} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_i-\theta)^2}{2}} =$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{(1-\theta)^2}{2}} \times \frac{1}{\sqrt{2\pi}} e^{-\frac{(0.5-\theta)^2}{2}} \times \frac{1}{\sqrt{2\pi}} e^{-\frac{(1.5-\theta)^2}{2}}$$

Using concrete numbers - 2

This is like maximizing:

$$e^{-\frac{(1-\theta)^2}{2}} \times e^{-\frac{(0.5-\theta)^2}{2}} \times e^{-\frac{(1.5-\theta)^2}{2}} =$$

$$= e^{-\frac{(1-\theta)^2}{2} - \frac{(0.5-\theta)^2}{2} - \frac{(1.5-\theta)^2}{2}} =$$

$$= e^{-\frac{(1-\theta)^2 + (0.5-\theta)^2 + (1.5-\theta)^2}{2}} = e^{-\frac{3.5 - 6\theta + 3\theta^2}{2}}$$

This is like minimizing $g(\theta) = 3.5 - 6\theta + 3\theta^2$.

$$\frac{dg(\theta)}{d\theta} = \frac{d3.5 - 6\theta + 3\theta^2}{d\theta} = -6 + 6\theta$$

Which becomes 0 for $\theta = 1$

What we have discovered

Our solution is therefore $\theta = 1$ and the desired pdf is $N(1, 1)$. But ...

$$mean(1, 0, 5, 1.5) = 1$$

We can try to generalize it...

Generalizing ...

Let's suppose to have:

- n iid random variables y_i with $i \in [1 \dots n]$
- with the same partially unknown pdf, that is
- $(\forall i)\ y_i \sim N(\theta, \sigma)$
- θ and σ to be determined.

We want to determine the value of θ that maximizes the probability of obtaining $(\forall i)\ y_i$.

In other terms our objective is to maximize the probability of occurrence of all y_i, that is a maximum likelihood estimation.

Typically, we would perform a least square estimation, and we know that optimal least square estimator is the Gaussian centered in the average of the points, with their standard deviation.

Maximum likelihood estimator (again)

Let' look for a maximum likelihood estimator!

$$\max_{\sigma,\theta}(\prod_{i=1}^{n} P(y_i|\sigma,\theta)) = \max_{\sigma,\theta}(\prod_{i=1}^{n} P(y_i|\sigma,\theta)) = \max_{\sigma,\theta}(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-\theta)^2}{2\sigma^2}}) =$$

$$= \max_{\sigma,\theta}\left((\frac{1}{\sigma\sqrt{2\pi}})^n \prod_{i=1}^{n} e^{-\frac{(y_i-\theta)^2}{2\sigma^2}}\right)) = \max_{\sigma,\theta}\left((\frac{1}{\sigma\sqrt{2\pi}})^n e^{-\sum_{i=1}^{n}\frac{(y_i-\theta)^2}{2\sigma^2}}\right) =$$

$$= \max_{\sigma,\theta}\left((\frac{1}{\sigma\sqrt{2\pi}})^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2}\right)$$

At this point we can take the log of the expression, knowing that the log function is differentiable and monotonically increasing on all $\mathbb{R}$.

Computing the ml estimator - 1

$$log\big((\frac{1}{\sigma\sqrt{2\pi}})^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2}\big) =$$

$$= n \times log(\frac{1}{\sigma\sqrt{2\pi}}) + log(e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2}) =$$

$$= n \times log(\frac{1}{\sigma\sqrt{2\pi}}) - \frac{1}{2\sigma^2} \times \sum_{i=1}^{n}(y_i-\theta)^2$$

Taking the partial derivative over θ we obtain:

$$\frac{\partial\Big(n \times log(\frac{1}{\sigma\sqrt{2\pi}}) - \frac{1}{2\sigma^2} \times \sum_{i=1}^{n}(y_i-\theta)^2\Big)}{\partial\theta} =$$

Computing the ml estimator - θ

$$= -\frac{\partial\left(\frac{1}{2\sigma^2} \times \sum_{i=1}^{n}(y_i - \theta)^2\right)}{\partial\theta} = -\frac{1}{\sigma^2} \times \left(\sum_{i=1}^{n}(y_i - \theta)\right)$$

And equating it to 0:

$$-\frac{1}{\sigma^2} \times \left(\sum_{i=1}^{n}(y_i - \theta)\right) = 0 \Rightarrow \sum_{i=1}^{n} y_i = n \times \theta \Rightarrow \theta = \frac{\sum_{i=1}^{n} y_i}{n}$$

Oh! θ is the average of the observed y_i!

Computing the ml estimator - σ - 1

$$\frac{\partial\left(n \times log(\frac{1}{\sigma\sqrt{2\pi}}) - \frac{1}{2\sigma^2} \times \sum_{i=1}^{n}(y_i - \theta)^2\right)}{\partial\sigma} =$$

$$= \frac{\partial\left(n \times log(\frac{1}{\sigma\sqrt{2\pi}})\right)}{\partial\sigma} - \frac{\partial\left(\frac{1}{2\sigma^2} \times \sum_{i=1}^{n}(y_i - \theta)^2\right)}{\partial\sigma} =$$

$$= -\frac{n}{\sigma} + \frac{1}{\sigma^3} \times \sum_{i=1}^{n}(y_i - \theta)^2$$

And equating it to 0:

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3} \times \sum_{i=1}^{n}(y_i - \theta)^2 = 0 \Rightarrow \left(\sum_{i=1}^{n}(y_i - \theta)^2\right) \times \frac{1}{\sigma^3} = \frac{n}{\sigma}$$

Computing the ml estimator - σ - 2

Assuming $\sigma \neq 0$:

$$\Rightarrow \left(\sum_{i=1}^{n} (y_i - \theta)^2 \right) = n \times \sigma^2 \Rightarrow$$

But we know $\theta = \overline{y_i}$, therefore:

$$\Rightarrow \sigma^2 = \frac{1}{n} \times \left(\sum_{i=1}^{n} (y_i - \overline{y_i})^2 \right)$$

Oh! σ is the standard deviation of the observed y_i!

What we have found

We have determined that the maximum likelihood estimator for a sequence of points assumed to be distributed normally is formed by a normal distribution with:

- average equal to the average of the sample,
- standard deviation equal to the standard derivation of the sample.

This coincides with the best quadratic estimator!

We now move forward considering the maximum likelihood estimator for a regression line, meaning, what happens if now we want to model an interdependencies using as objective function the maximum likelihood.

Ml linear regression - HPs

Let's suppose to have:

- $n \times m$ values $x_{i,j}$ with $i \in [1 \dots n]$, $j \in [1 \dots m]$ represented in short by a matrix $\boldsymbol{X}$ or a vector $\mathbf{x_i}$, $n > m$ *(why?)*
- n iid random variables y_i with $i \in [1 \dots n]$ represented in short by a vector $\boldsymbol{y}$
- a linear relationships $\boldsymbol{\theta}$ between $\boldsymbol{X}$ and $\boldsymbol{y}$, *that is, we use the usual link / mean functions*
- each y_i distributed normally with mean $\boldsymbol{x_i^T \theta}$ and standard deviation σ (the same σ for all y_i), that is
- $(\forall i)\ y_i \sim N(\boldsymbol{x_i^T \theta}, \sigma)$
- $\boldsymbol{\theta}$ and σ to be determined.

Ml linear regression - goals

We want to determine the value of $\boldsymbol{\theta}$ and σ that maximizes the probability of obtaining $(\forall i)\ y_i$, that is:

$$\max_{\theta,\sigma}(P(\mathbf{y}|\boldsymbol{X},\boldsymbol{\theta},\sigma)) = \max_{\boldsymbol{\theta},\sigma}(\prod_{i=1}^{n} P(y_i|\boldsymbol{x_i},\boldsymbol{\theta},\sigma))$$

In other terms, our objective function is the conditional probability of occurrence of all y_i.

Computing the optimal θ - 1

We can express for simplicity our equation in vectorial form:

$$\max_{\sigma,\boldsymbol{\theta}} \left(\left(\frac{1}{\sigma\sqrt{2\pi}} \right)^n e^{-\frac{(\boldsymbol{y}-\boldsymbol{X\theta})^T(\boldsymbol{y}-\boldsymbol{X\theta})}{2\sigma^2}} \right)$$

As mentioned, this is equivalent to maximizing the log:

$$\max_{\sigma,\boldsymbol{\theta}} \left(log \left(\left(\frac{1}{\sigma\sqrt{2\pi}} \right)^n e^{-\frac{(\boldsymbol{y}-\boldsymbol{X\theta})^T(\boldsymbol{y}-\boldsymbol{X\theta})}{2\sigma^2}} \right) \right)$$

Which becomes:

$$\max_{\sigma,\boldsymbol{\theta}} \left(n \times log \left(\frac{1}{\sigma\sqrt{2\pi}} \right) + log \left(e^{-\frac{(\boldsymbol{y}-\boldsymbol{X\theta})^T(\boldsymbol{y}-\boldsymbol{X\theta})}{2\sigma^2}} \right) \right)$$

Computing the optimal θ - 2

$$\max_{\sigma,\boldsymbol{\theta}} \left(n \times log\left(\frac{1}{\sqrt{2\pi}}\right) + n \times log\left(\frac{1}{\sigma}\right) - \frac{(\boldsymbol{y}-\boldsymbol{X\theta})^T(\boldsymbol{y}-\boldsymbol{X\theta})}{2\sigma^2} \right)$$

And now we take the partial derivative over $\boldsymbol{\theta}$:

$$\frac{\partial \left(n \times log\left(\frac{1}{\sqrt{2\pi}}\right) + n \times log\left(\frac{1}{\sigma}\right) - \frac{(\boldsymbol{y}-\boldsymbol{X\theta})^T(\boldsymbol{y}-\boldsymbol{X\theta})}{2\sigma^2} \right)}{\partial \boldsymbol{\theta}} =$$

$$= -\frac{1}{2\sigma^2} \frac{\partial \left((\boldsymbol{y}-\boldsymbol{X\theta})^T(\boldsymbol{y}-\boldsymbol{X\theta}) \right)}{\partial \boldsymbol{\theta}} = -\frac{1}{\sigma^2}(\boldsymbol{y}-\boldsymbol{X\theta})$$

And equating it to 0 we obtain:

$$-\frac{1}{\sigma^2}(\boldsymbol{y}-\boldsymbol{X\theta}) = 0 \quad \Rightarrow \quad \boldsymbol{y} = \boldsymbol{X\theta}$$

Computing the optimal θ - 3

If $\boldsymbol{X}$ were square, then the solution would be:

$$\boldsymbol{\theta} = \boldsymbol{X}^{-1}\boldsymbol{y}$$

But, as we said, $n > m$, therefore the solution is given by:

$$\boldsymbol{\theta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

What a surprise, isn't it?

Computing the optimal σ

Starting from:

$$\max_{\sigma,\boldsymbol{\theta}} \left(n \times log\left(\frac{1}{\sqrt{2\pi}}\right) + n \times log\left(\frac{1}{\sigma}\right) - \frac{(\boldsymbol{y}-\boldsymbol{X\theta})^T(\boldsymbol{y}-\boldsymbol{X\theta})}{2\sigma^2} \right)$$

And now we take the partial derivative over σ:

$$\frac{\partial \left(n \times log\left(\frac{1}{\sqrt{2\pi}}\right) + n \times log\left(\frac{1}{\sigma}\right) - \frac{(\boldsymbol{y}-\boldsymbol{X\theta})^T(\boldsymbol{y}-\boldsymbol{X\theta})}{2\sigma^2} \right)}{\partial\sigma} =$$

$$= -\frac{n}{\sigma} + \frac{(\boldsymbol{y}-\boldsymbol{X\theta})^T(\boldsymbol{y}-\boldsymbol{X\theta})}{\sigma^3}$$

And equating it to 0, assuming as usual $\sigma \neq 0$ we obtain:

$$\sigma^2 = \frac{(\boldsymbol{y}-\boldsymbol{X\theta})^T(\boldsymbol{y}-\boldsymbol{X\theta})}{n}$$

Maximum likelihood estimator - properties

Claim 1: The maximum likelihood estimator of a Gaussian distribution over a set of points coincides with the OLS estimator.
Proof: See above.
QED

Claim 2: The maximum likelihood linear regression coincides with the OLS linear regression.
Proof: See above.
QED

Bernoulli and maximum likelihood

The pdf of a Bernulli distribution can be represented in terms of conditional probability as:

$$P(x|\theta) = \theta^x(1-\theta)^{(1-x)}$$

where clearly x can only be 0 or 1.

We can now introduce the concept of entropy, already hinted in class. Entropy represents the level of uncertainty of a variable.

Entropy (and Bernulli and ml)

Definition (Entropy): Given a random vectorial variable x of n components and a parameter θ, we define entropy of x, $H(x)$ as:

$$H(x) = \sum_{i=1}^{n} p(x_i|\theta) \times log(p(x_i|\theta))$$

We notice that for a Bernulli distribution:

$$H(x) = (1-\theta)log(1-\theta) + \theta log(\theta)$$

Indeed, as θ tends to 0 or to 1 the uncertainty tends to 0, since the likely value of x tend to be 0 or 1 respectively.

From B&B plus ml to LR

We are now ready to move to study a radically different form or regression, the so-called logistic regression.

Our goal is to have a regression that not only represents a relationship between two variables, but is also possible to capture a prediction of probability.

However, the value of a probability is from 0 to 1, so we need a "good" function that can translate any value in such range.

We use often as such function the so-called "sigmoid function." To introduce the sigmoid we start with the definition of a "logistic function."

Logistic

Definition (Logistic function): Given $L, x_0 \in \mathbb{R}$, $k \in \mathbb{R}^+$ a logistic function $f(x)$ is defined as:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

Properties (of the logistic function:)

- the domain is all $\mathbb{R}$
- the range is $[0 \dots L]$ if L is positive and $[L \dots 0]$ if L is negative
- $f(x)$ is continuous, monotonically increasing, and differentiable over all its domain
- $f(x)$ is symmetric over x_0
- k is the rate of growth of $f(x)$ and for $k \to +\infty$ $f(x)$ tends to become the step function in x_0

Sigmoid

Definition (Sigmoid): Given $k \in \mathbb{R}^+$, a sigmoid function $sigm(x)$ is defined as a logistic function with $L = 1$ and $x_0 = 0$:

$$sigm(x) = \frac{1}{1 + e^{-kx}}$$

Properties (of the sigmoid function):

- the domain is all $\mathbb{R}$
- the range is $[0 \dots 1]$
- $sigm(x)$ is continuous, monotonically increasing, and differentiable over all its domain
- $sigm(x)$ is symmetric over 0
- k is the rate of growth of $sigm(x)$ and for $k \to +\infty$ $sigm(x)$ tends to become the step function

Toward a logistic regression

Suppose that we want to determine if a given event is going to happen based on a series of n predictors $x_1 \dots x_n$. We can model the probability of occurrence of the event with a random variable y.

It is as if we have a sequence of flipping of coins each with different values of the possible variables that affect the result, for instance the intensity of the flipping, the temperature, the wind, etc.

Based on such set we want to predict what will be the result of the next flipping, given a set of values assigned to the covariates.

Our question is what is:

$$\text{P(Head} \mid \text{strong toss, strong wind, 60 degrees)}$$

?

Toward a logistic regression

Let's try to build a regression line.

As we mentioned, any time we compute a regression we need to determine:

- the function to use as a model, and in this case a linear function would not be suitable, since probabilities range from 0 to 1, for this reason we select a sigmoid function;
- the objective function, and in this case the least square would be inappropriate because it is not a proper metrics space, so we opt for maximizing the conditional probability, that is, we aim at a maximum likelihood estimation.

Logistic regression - HPs

Let

- (y_i, x_i) be a collection of pairs with:
  - $i \in [1 \dots n]$
  - $y_i \in \{0, 1\}$
  - $x_i \in \mathbb{R}^m$
  - $n > m$
- assume that the y_i are iid random variables
- consider as target mean function the sigmoid
- consider as optimality criteria the maximum likelihood

Logistic regression - goals

We want to determine the values of the parameters that maximize the probability of obtaining $(\forall i)\ y_i$, that is:

$$\max_{Parameters}(P(\mathbf{y}|\boldsymbol{X}, Parameters) = \max_{\boldsymbol{\theta}}(\prod_{i=1}^{n} P(y_i|\boldsymbol{x_i}, Parameters))$$

In other terms, our objective function is the conditional probability of occurrence of all y_i.

Given our link/mean:

$$\max_{\theta}(P(\mathbf{y}|\boldsymbol{X}, \boldsymbol{\theta})) = \max_{\boldsymbol{\theta}}(\prod_{i=1}^{n} P(y_i|sigm(\boldsymbol{x_i}^T\boldsymbol{\theta}))$$

Logistic regression - structure

Since the pdf of a Bernulli distribution is:

$$P(z|k) = k^z(1-k)^{(1-z)}$$

For us the probability k of each event is "approximated" by the sigmoid function (our mean function):

$$\boldsymbol{k} = \frac{1}{1 + e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}$$

And this lead us to

$$P(y_i|\boldsymbol{x_i}, \boldsymbol{\theta}) = \left(\frac{1}{1 + e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}\right)^{y_i} \times \left(1 - \frac{1}{1 + e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}\right)^{1-y_i}$$

Logistic regression - the problem

Our problem has therefore the form of:

$$\max_{\theta}(P(\mathbf{y}|\boldsymbol{X},\boldsymbol{\theta})) = \max_{\boldsymbol{\theta}} \prod_{i=1}^{n} \left(\frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}\right)^{y_i} \times \left(1-\frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}\right)^{1-y_i}$$

It is like finding an n-dimensional hyperplane dividing the n-dimensional hyperspace in 2 parts, those leading to y being 0 and those leading to y being 1.

Logistic regression - solution - 1

Since the log function is continuous, differentiable and monotonically increasing in all $\mathbb{R}^+$, our problem is equivalent to:

$$\max_{\boldsymbol{\theta}} \left(log \left(\prod_{i=1}^{n} \left(\frac{1}{1+e^{-\boldsymbol{x_i}^T \boldsymbol{\theta}}} \right)^{y_i} \times \left(1 - \frac{1}{1+e^{-\boldsymbol{x_i}^T \boldsymbol{\theta}}} \right)^{1-y_i} \right) \right)$$

And, given the property of logs, this is like maximizing:

$$log \left(\prod_{i=1}^{n} \left(\frac{1}{1+e^{-\boldsymbol{x_i}^T \boldsymbol{\theta}}} \right)^{y_i} \right) + log \left(\prod_{i=1}^{n} \left(1 - \frac{1}{1+e^{-\boldsymbol{x_i}^T \boldsymbol{\theta}}} \right)^{1-y_i} \right) =$$

$$= \sum_{i=1}^{n} log \left(\frac{1}{1+e^{-\boldsymbol{x_i}^T \boldsymbol{\theta}}} \right)^{y_i} + \sum_{i=1}^{n} log \left(1 - \frac{1}{1+e^{-\boldsymbol{x_i}^T \boldsymbol{\theta}}} \right)^{1-y_i} =$$

Logistic regression - solution - 2

$$= \sum_{i=1}^{n} y_i \times log\left(\frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}\right) + \sum_{i=1}^{n}(1-y_i) \times log\left(1-\frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}\right) = \cdots$$

A bit of logarithms...

$$log\left(\frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}\right) = log(1) - log\left(1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}\right) = -log\left(1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}\right)$$

$$log\left(1-\frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}\right) = log\left(\frac{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}-1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}\right) = log\left(\frac{e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}\right) =$$

$$= log\left(e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}\right) - log\left(1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}\right) = \boldsymbol{x_i}^T\boldsymbol{\theta} - log\left(1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}\right) =$$

Logistic regression - solution - 3

$$= -\sum_{i=1}^{n} y_i \times log\left(1 + e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}\right) + \sum_{i=1}^{n}(1-y_i) \times \left(\boldsymbol{x_i}^T\boldsymbol{\theta} - log\left(1 + e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}\right)\right) =$$

For simplicity let w_i be $log\left(1 + e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}\right)$.

$$= -\sum_{i=1}^{n} y_i \times w_i + \sum_{i=1}^{n} \boldsymbol{x_i}^T\boldsymbol{\theta} - \sum_{i=1}^{n} w_i - \sum_{i=1}^{n} y_i \times \boldsymbol{x_i}^T\boldsymbol{\theta} + \sum_{i=1}^{n} y_i \times w_i =$$

$$= \sum_{i=1}^{n} \boldsymbol{x_i}^T\boldsymbol{\theta} - \sum_{i=1}^{n} w_i - \sum_{i=1}^{n} y_i \times \boldsymbol{x_i}^T\boldsymbol{\theta} =$$

$$= \sum_{i=1}^{n}(1 + y_i)\boldsymbol{x_i}^T\boldsymbol{\theta} - \sum_{i=1}^{n} w_i$$

Logistic regression - comments

Let $f(\theta)$ be:

$$\sum_{i=1}^{n}(1+y_i)\boldsymbol{x_i}^T\boldsymbol{\theta} - \sum_{i=1}^{n} log\left(1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}\right)$$

Claim: $f(\theta)$ is convex.

Proof: Omitted

Consequence: Optimization algorithms can easily find the maximum.