Week4

Saturday, February 26, 2022   12:32 PM

Logistic   Regression

→ This a classification technique
→ Distribution convergence: when they both reach the same distribution

win

Outcome
0
Distance from basket
→ result

the shorter the distances, the more likely a win is
→ at what distance should we shoot from?
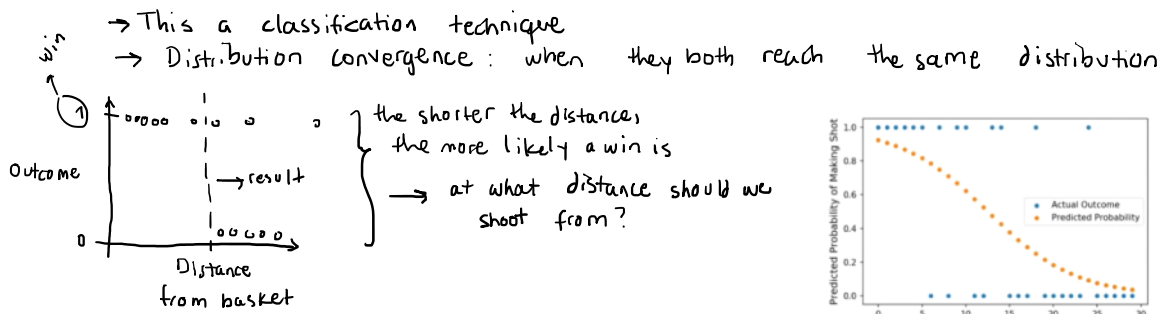
Verisimilitude (verosimilitud)
↳ how similar my data is, called π

0.7
0.1
evaluated → verisimilitude is the fact that we have an x and we evaluate on the same points (the PDF function)

model
than this one
the model is better in this area
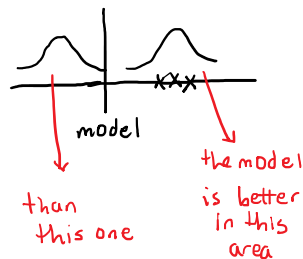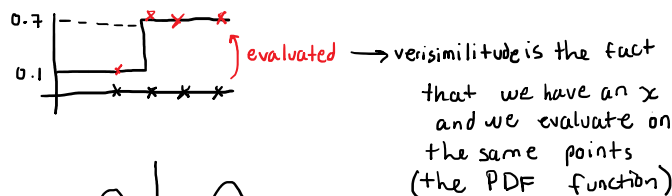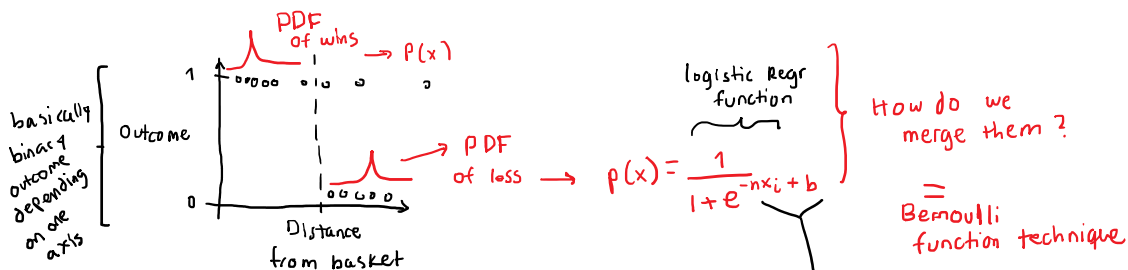
the more x points land on the distribution's density domain, the larger π we get, the better model
→ (probability density function
look for the closest points to the density function's peak.

we want to maximize it
$$\pi = (0.1)(0.7)(0.7) = \boxed{0.049}$$

PDF of wins → P(x)

1
Outcome
0
Distance from basket

basically binary outcome depending on one axis

PDF of loss → $p(x) = \dfrac{1}{1+e^{-nx_i + b}}$

logistic Regr function

How do we merge them?
=
Bernoulli function technique

filters what we consider when merging the PDF's for both win/loss

Bernoulli function : $f(x) = p(x_i)^{y_i}(1-p(x_i))^{1-y_i}$
(Bernoulli Trials)

x is verisimilitude

we want to find the m, b parameters that optimize the function

Thus

$$\max_{m/b} f(x) = \max_{m,b} \prod_{i=\emptyset}^{N} f(x_i)$$

prod

$$= \max_{m,b} \prod_{i=1}^{N} p(x_i)^{y_i}(1-p(x_i))^{1-y_i}$$

After log( ), $\log(ab) = \log(a) + \log(b)$

$$= \max_{m,b} \log\left(\prod p(x_i)^{y_i}(1-p(x_i))^{1-y_i}\right)$$

+ now Σ

$$= \max_{m,b} \sum_{i=0}^{N-1} \log\left(p(x_i)^{y_i}(1-p(x_i))^{1-y_i}\right)$$

Why $\log(L(x_i, m, b))$?

Y
max
max

Same number

log(x)
↳ monotonously incrementing
=
do not affect the maximum

since the increment softens
$\log(x)$ "softens the curve" and reduces dimensions
↓
useful for maximizing

f(x)
max (same point)
log(x)

Differentiate to maximize (and also equalize to ∅)
$$\frac{d}{dm} \sum_{i=0}^{N-1} \log\left(p(x_i)^{y_i}(1-p(x_i))^{1-y_i}\right) = 0$$

$$\rightarrow \sum_{i=0}^{N-1} \dots$$

$$\frac{d}{dm} \sum_{i=0} \log\left(p(x_i)^{y_i} (1-p(x_i))^{\cdot y_i} \right) = 0$$

$$\frac{d}{dm} \sum_{i=0}^{N-1} \log\left(p(x_i)^{y_i} \right) + \log\left(1-p(x_i) \right)^{1-y_i} = 0$$

$$\frac{d}{dm} \sum_{i=0}^{N-1} y_i \log\left(p(x_i) \right) + (1-y_i) \log\left(1-p(x_i) \right) = 0$$
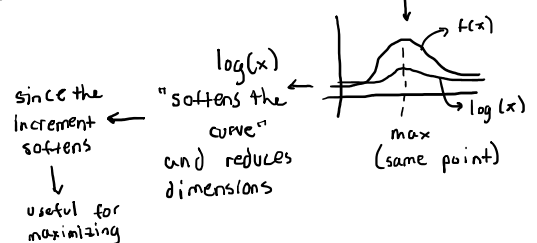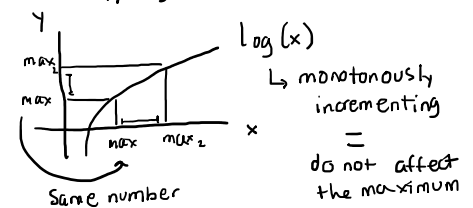
$$\frac{d}{dm} \sum_{i=0}^{N-1} \left(y_i \log\left(p(x_i) \right) + \log\left(1-p(x_i) \right) - y_i \log\left(1-p(x_i) \right) \right) = 0$$

useful for maximizing

$\boxed{x_i = \text{distance to the basket} \\ y_i = 0 \text{ or } 1 \ (\text{fail or success})}$

Considering $\quad P(x_i) = \frac{1}{1+e^{-mx_i+b}}$, with bias $b=0$

$\underbrace{}_{\text{logistic distribution}}$

$$\frac{d}{dm} \sum \left[y_i \log\left(\frac{P(x_i)}{1-P(x_i)} \right) + \log\left(1 - P(x_i) \right) \right] = 0$$

Evaluating the term $\frac{P(x_i)}{1-P(x_i)}$, considering $\alpha = mx_i + b$

$$\frac{P(x_i)}{1-P(x_i)} = \frac{\frac{1}{1+e^{-\alpha}}}{1 - \frac{1}{1+e^{-\alpha}}} = \frac{\frac{1}{1+e^{-\alpha}}}{\frac{\cancel{1}+e^{-\alpha}-\cancel{1}}{1+e^{-\alpha}}} = \frac{1+e^{-\alpha}}{e^{-\alpha}(1+e^{-\alpha})} = \frac{1}{e^{-\alpha}} = e^{\alpha}$$

$$= e^{mx_i+b}$$

Evaluating the term $1 - P(x_i)$

$$1 - P(x_i) = 1 - \frac{1}{1+e^{-\alpha}} = \frac{\cancel{1}+e^{-\alpha}-\cancel{1}}{1+e^{-\alpha}} = \frac{e^{-\alpha}}{1+e^{-\alpha}}$$

Thus,

$$= \frac{d}{dm} \sum_{i=0}^{N-1} y_i \log\left(e^{mx_i+b} \right) + \sum_{i=0}^{N-1} \log\left(\frac{e^{-(mx_i+b)}}{1+e^{-(mx_i+b)}} \right) = 0$$

Now with b=0 (bias)

$$\frac{d}{dm} \left(-\sum_{i=0}^{N-1} mx_i y_i - \sum_{i=0}^{N-1} mx_i + \sum_{i=0}^{N-1} \log\left(1+e^{-mx_i} \right) \right) = \frac{d}{dm} \ell(x,m) = 0$$

$\underbrace{}_{\max\limits_{m} \ell(x,m)}$

Differentiating

$$= 2 \sum_{i=0}^{N-1} \left(x_i y_i + \frac{x_i(1+e^{-mx})}{1+e^{-mx}} \right)$$

$$= \boxed{\sum_{i=0}^{N-1} x_i (y_i - P(x_i))}$$

$\downarrow$

this sum will output the m that maximizes the verisimilitude function

where the objective was to find parameters m and b that maximize the verisimilitude function

$$\max_{m,b} L(x_i, m, b) = \max_{m,b} \prod f(x_i)$$

Types of Convergence

- Convergence in distribution, converges weakly or in law (fractals)

Convergence in mean (unbiased estimator)

- Convergence 'Almost - Sure':
  - Almost surely, almost everywhere, strongly or with probability 1.

We thus said we try to find m,b parameters that maximize our function, that is, that separate the training data in almost-sure convergence. The problem is addressed with the Likelihood Method, which is defined for Bernoulli Trials as mentioned.

$$L(m,b) = \prod_{i=0}^{N-1} p(x_i)^{y_i} (1-p(x_i))^{1-y_i}$$

$\longrightarrow$ Thus $\max\limits_{m,b} L(m,b)$

Our results for maximizing the Likelihood Function (verisimilitude) were

$$\sum_{i=0}^{N-1} \left(\cdots \right) \cdots$$

Our results for maximizing the Likelihood Function (verisimilitude) were

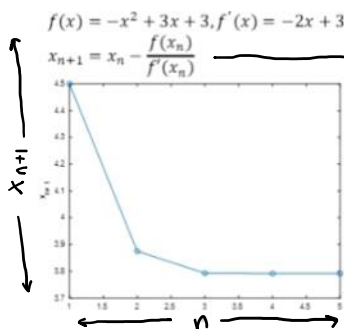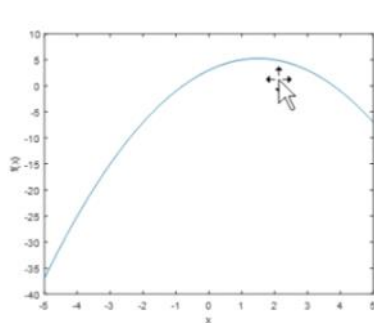$$\frac{\partial \ell(m,b)}{\partial m} = \sum_{i=0}^{N-1} \left((y_i - p(x_i)) x_i \right) = 0$$

But unfortunately the last equation does not present a closed form, and thus we propose to use a Newton-Rapson method for numerical approximation

$$m_{iteration+1} = m_{iteration} - \frac{\frac{\partial \ell(m_{iteration})}{\partial m}}{\frac{\partial^2 \ell(m_{iteration})}{\partial^2 m}}$$

Basically we can't solve this for m since m is very inside of an exponential function, so we will use a num method for 'solving' for m.

similar

Newton Rapson Example
• Given the f(x) function, find its maximum

$$f(x) = -x^2 + 3x + 3, f'(x) = -2x + 3$$
$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

x_{n+1}

n

So, our expression of Newton Rapson to the Likelihood Function

$$m_{iteration+1} = m_{iteration} - \frac{\sum_{i=0}^{N-1} \left((y_i - p(x_i) x_i \right)}{\sum_{i=0}^{N-1} \left(-x_i(1 - p(x_i)) - p(x_i) \right)}$$

Computing $\frac{\frac{\partial \ell}{\partial m}}{\frac{\partial^2 \ell}{\partial^2 m}}$

Example,
    Given a set of training, find the parameter m and b such that separates into classes

$$X = \{-0.1, -0.5, -2, 2, 1.3, 4, 0, 5\}$$
$$Y = \{0\ 0\ 0\ 1\ 1\ 1\ 1\ 1\}$$

    — The stop criteria between iterations is > 0.05

    → Code Available, resulting in:

b = 0 (bias)
m = 1.2140
error = -0.043547