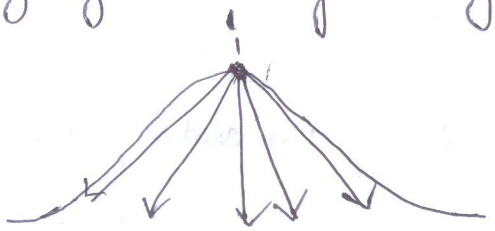
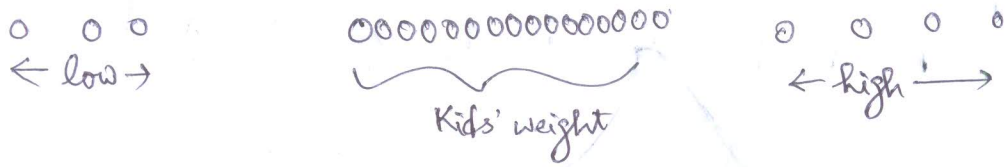
MLE & MAP (ML Lecture 11)

Experiment: weighing a bunch of nursery kids





Goal: MLE fits distribution to data, in a maximum likely way. There exist diff distributions (normal, exponential, gamma & many more ...)

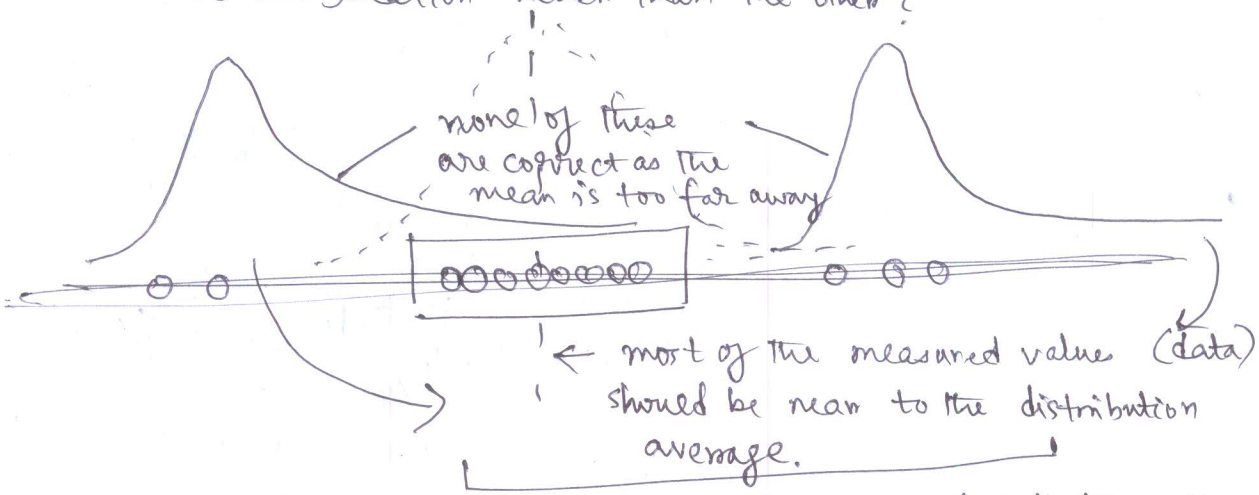
⇒ convenient !

In this case, we assume the weights to be normally distributed. This implies

i) most of the measurements to be close to ~~zero~~ the mean

ii) most of the measurements relatively symmetrical to the mean.

Once we settle on the shape, we have to settle the location of the distribution. Is one location better than the other ?

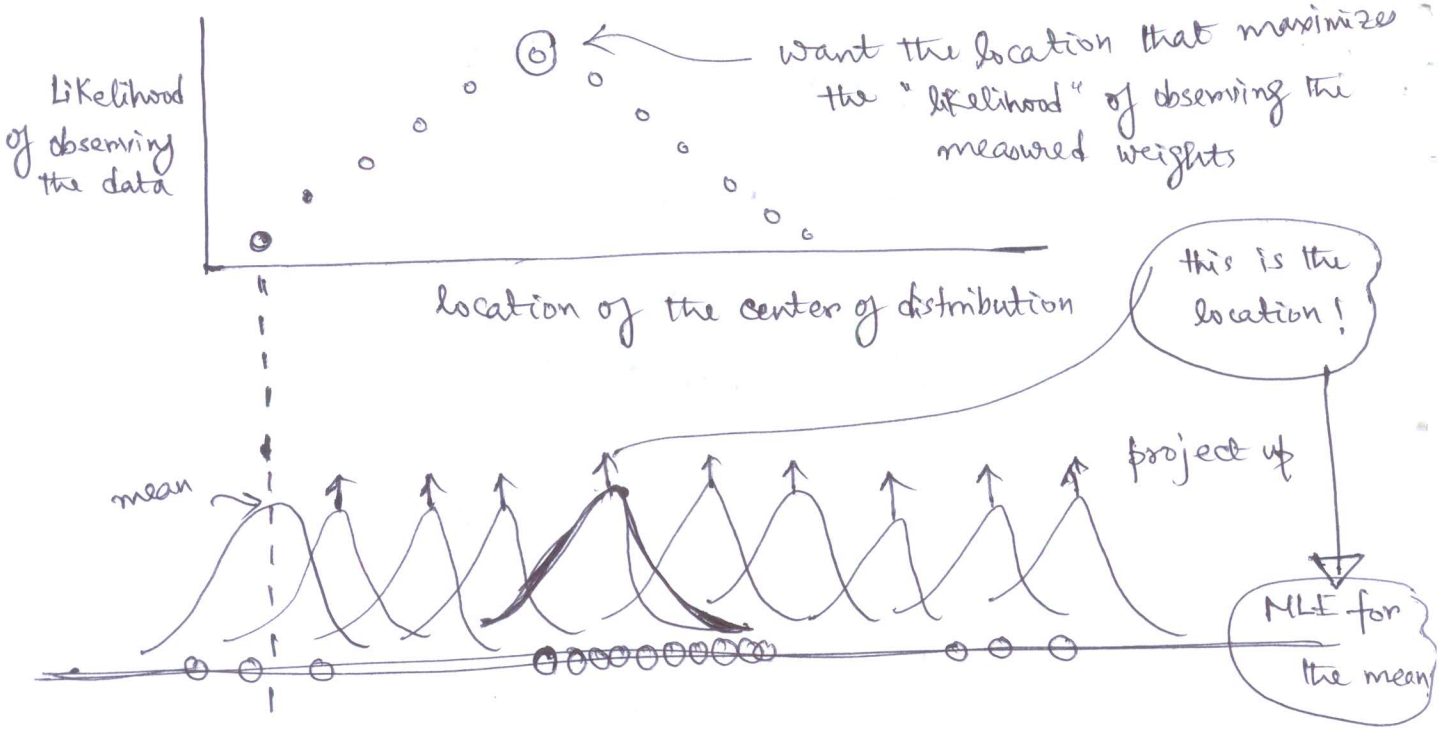


not satisfied by the distribution on the far left/right.

So, what does it mean? ⟸

↓

the likelihood of observing all the weights is low; however for the distribution at the center, the likelihood of observing most measurements is high.



Similarly, we compute the MLE for the standard deviation.

We have the MLE estimates for the mean ⟺ we found the value for mean OR SD that maximizes the likelihood that we observed the data

MLE tries to find the optimal value for the mean/SD for a distribution given a bunch of measurements

MLE for the Normal distribution:

$$P(x/\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

(location of mean, location of SD)

likelihood function → $P(D/\theta) \Leftrightarrow P(\mu,\sigma/x)$

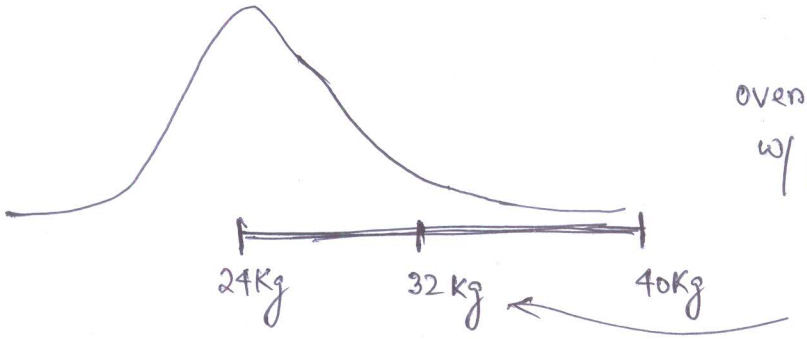
Location of μ: smaller μ value → moves the distribution to the left

larger μ value → ——— right

Location of σ: Determines the width of the distribution

smaller σ → distribution taller & narrower (lean)

larger σ → distribution fatter & wider (bored)



overlay a normal distribution w/ $\mu = 28$ & $\sigma = 2$ onto the data

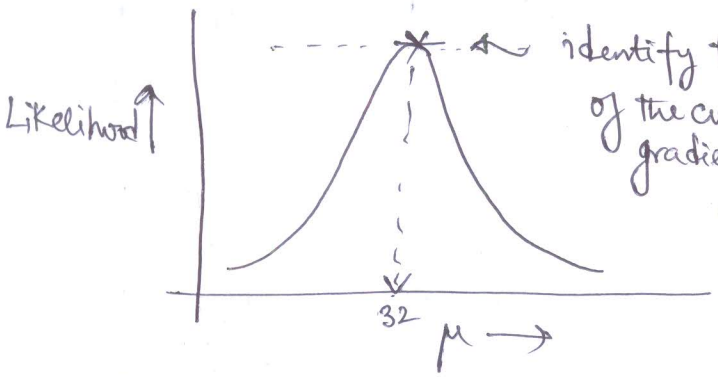
single measurement of weight

To determine the likelihood of the data given this distribution, we can plug the numbers in the likelihood function: $\longrightarrow L(\mu = 28, \sigma = 2 / x = 32)$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} = \frac{1}{\sqrt{2\pi \cdot 2^2}} e^{-(32-28)^2/2 \cdot 2^2} = 0.03$$

Now, shift the distribution to the right i.e $\mu = 30. \Rightarrow L = 0.12$

Do this for a bunch of different distributions $(\mu = 20) \rightarrow L() = 0.0000\ldots0003$ really small



Fix 32 for μ & vary σ for the appropriate distribution $\Rightarrow$ need to do this for multiple data points. Therefore,

MLE for μ: σ is constant, find the μ for which slope = 0

MLE for σ: μ is constant, — σ — slope = 0

More data:

$x_1 = 32$ Kg
$x_2 = 34$ Kg

L_1: $L(\mu = 28, \sigma = 2 / x_1 = 32)$; L_2: $L(\mu = 28, \sigma = 2 / x_2 = 34)$

$x_1, x_2 \in D$

But what's the likelihood when $x_1 = 32$ & $x_2 = 34$ i.e

$$L(\mu = 28, \sigma = 2 / x_1 = 32 \ \& \ x_2 = 34) = L_1 * L_2$$

$L(\theta / D)$ should be written as $L(D/\theta)$

$\Longrightarrow$ IID (measurements are independent, weighing x_1 doesn't have an effect on weighing x_2)

Keep adding data points ; i.e $D = \{x_1, \dots x_n\}$ then,

$$L(\theta/D) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/2\sigma^2}$$

$$\underbrace{L = (\mu, \sigma / x_1, \dots x_n)}_{L(\theta/D)} = L(\mu,\sigma/x_1)\, L(\mu,\sigma/x_2) \dots L(\mu,\sigma/x_n)$$

MLE estimation
- derivative wrt μ (σ Constant)
- derivative wrt σ (μ Constant)

$$\ln L(\theta/D) = \ln\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/2\sigma^2}\right)$$

$$= \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_1-\mu)^2/2\sigma^2}\right) + \dots \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_n-\mu)^2/2\sigma^2}\right)$$

$$= -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(x_1-\mu)^2}{2\sigma^2}$$

$$= -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\cdot 2\ln(\sigma) - \frac{(x_1-\mu)^2}{2\sigma^2} + \dots$$

$$= -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$\frac{\partial L}{\partial \mu} = 0 \Rightarrow \sum_{i=1}^{n} + \frac{2(x_i-\mu)(+1)}{2\sigma^2} \Rightarrow \sum_{i=1}^{n} \frac{x_i-\mu}{\sigma^2} = 0 \Rightarrow \mu = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

($\bar{x}$: Sample mean)

$$\frac{\partial L}{\partial \sigma} = 0 \Rightarrow -\frac{n}{\sigma} + \sum_{i=1}^{n} (x_i-\mu)^2 (+2)\sigma^{-3} = 0$$

$$\Rightarrow -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^3} = 0$$

$$\Rightarrow \sigma^2 = \frac{\sum_{i=1}^{n} (x_i-\mu)^2}{n} \rightarrow \text{SD of measurements}$$

$$\Rightarrow \sigma = \left(\sum_{i=1}^{n} (x_i-\mu)^2/n\right)^{1/2}$$

MAP: Given, $D = \{x_1, \dots x_n \mid x_i \in R^n\}$; assume a joint distribution $p(D,\theta)$

θ is R.V ; Goal: choose a "good" θ for D (hypothesis / class label)

$$\theta_{MAP} = \underset{\theta}{\text{argmax}}\ p(\theta/D) \rightarrow \text{posterior density}$$

Compare w/ MLE

$$\theta_{MLE} = \underset{\theta}{\text{argmax}}\ p(D/\theta) \rightarrow \text{likelihood function}$$

Pros: (i) easy to compute & interpret ; $p(D,\theta) = p(D/\theta)\,p(\theta)$

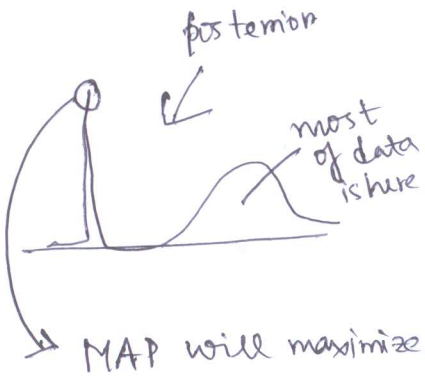
MAP interpolates b/w MLE & prior $\Leftarrow$ MLE, prior

(ii) avoids overfitting $\rightarrow$ MLE does

~~(iii)~~

Cons: (i) point estimate $\Rightarrow$ no representation of uncertainty

(ii) must assume prior on θ



MAP for univariate Gaussian: (mean)

$D = \{x_1, \dots x_n\}$, $\theta \rightarrow$ R.V $\sim N(\mu, 1)$ — prior mean

$\{x_1, \dots x_n\}$ IID

$$p(x_1, \dots x_n \mid \theta) = \prod_{i=1}^{n} p(x_i/\theta)$$

no data ⇒ θ is used in place of x_i's

$$\theta_{MAP} = \underset{\theta}{\text{argmax}}\ p(\theta/D) = \frac{p(D/\theta)\,p(\theta)}{p(D)} = \underset{\theta}{\text{argmax}}\ (p(D/\theta)\,p(\theta))$$

$p(D)$: Bayes Rule

$$= \underset{\theta}{\text{argmax}}\ \ln(p(D/\theta)\,p(\theta))$$

$$= \underset{\theta}{\text{argmax}}\ (\ln p(D/\theta) + \ln p(\theta))$$

MLE for $\theta \sim N(\mu, 1)$

$$\frac{\partial}{\partial\theta}(\ln p(D/\theta) + \ln p(\theta))$$

MLE for μ: $\sum_{i=1}^{n} \frac{(x_i - \mu)}{\sigma^2} = 0$

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

$$\frac{\partial}{\partial\theta} \ln p(\theta) = \frac{\partial}{\partial\theta} \ln\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\theta-\mu)^2}\right)$$

$$= \frac{\partial}{\partial\theta}\left(-\frac{1}{2}\ln 2\pi - \frac{1}{2}(\theta-\mu)^2\right)$$

$$= (\theta - \mu) \cdot -\theta$$

$$\therefore \frac{\partial}{\partial \theta}(\quad) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - n\theta\right) + (\mu - \theta) = 0$$

$$\Rightarrow \left(\frac{n}{\sigma^2} + 1\right)\theta = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i + \mu$$

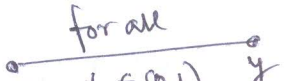
$\bar{x}$ ← sample mean

$$\Rightarrow \theta = \frac{\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i + \mu}{\frac{n}{\sigma^2} + 1} = \frac{n.\frac{1}{n}\sum x_i + \mu\sigma^2}{n + \sigma^2}$$

$$\boxed{\theta_{MAP} = \frac{n}{n+\sigma^2}\bar{x} + \frac{\sigma^2}{n+\sigma^2}\mu} \quad \text{---------- (*)}$$

Convex combination of sample mean μ & prior mean, $\bar{x}$ (from prior data)

Convex combination of x & y : $\alpha x + (1-\alpha) y$; $\alpha \in (0,1)$,



$\alpha = 0$, y

$\alpha = 1$, x

$\therefore \theta_{MAP} \sim \bar{x}$ when $\sigma^2 = 0$ (No variance)

(x_{MLE}) $\Rightarrow$ MLE is a special case of MAP

(i) As n & σ^2 vary, we obtain a range of values b/w sample mean ($\bar{x}$) & prior mean (μ)

(ii) $n = 0$ (No data), $\theta_{MAP} \sim \mu$ (prior mean)

(iii) $n \to \infty$, $\theta_{MAP} \sim \bar{x}$ (sample mean)

Large # of data, MAP estimator is the sample mean $\Leftrightarrow$ MLE

MAP: An illustration

$$P(\theta/D) = \frac{P(D/\theta)\,P(\theta)}{P(D)}$$

likelihood; prior; posterior

Recall the "cancer test" problem. What is MAP hypothesis?

Recall;

(i) cancer / + class; ¬cancer → − class; Assume $\theta \equiv h$

(ii) Confusion matrix:

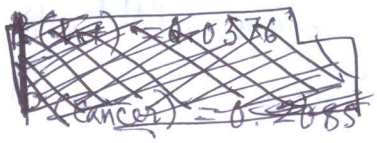
Actual \ Prediction	−	+
−	0.97	0.03
+	0.02	0.98

$$P(h/D) = \frac{P(D/h)\,P(h)}{P(D)}$$

(iii)

$P(h-) = 0.992$
$P(h+) = 0.008$

$P(- | \neg\text{cancer}) = 0.97$
$P(- | \text{cancer}) = 0.02$
$P(+ | \neg\text{cancer}) = 0.03$
$P(+ | \text{cancer}) = 0.98$



Now,

$$h_{MAP} = \underset{h}{\text{argmax}}\; P(D/h)\,P(h)$$

$P(+|\text{cancer})\,P(\text{cancer}) = 0.98 \times .008 = 0.00784 = P(h+/D)$ ← prob(cancer / tested positive)

$P(+|\neg\text{cancer})\,P(\neg\text{cancer}) = 0.03 * 0.992 = 0.02976 = P(h-/D)$ ← prob(¬cancer / tested positive)

$$\boxed{h_{MAP} = 0.02976}$$

Given data, D; choose a good hypothesis / class label for D.

$h+ \equiv$ Cancer
$h- \equiv \neg$ cancer
} two hypotheses;

For each hypothesis in H, compute the posterior probability

$$P(h+/D) = \frac{P(D/h+)\,P(h+)}{P(D)} = 0.00784 \quad ; \quad P(h-/D) = \frac{P(D/h-)\,P(h-)}{P(D)} = 0.02976$$

Output the hypothesis h_{MAP} with the highest posterior probability

$$h_{MAP} = \underset{h \in H}{\text{argmax}} \{P(h+/D), P(h-/D)\} = P(h-/D)$$

Note on MLE & MAP:

1. MAP can be interpreted as forming & changing opinion as an evolutionary process. You have an apriori belief that 'decides in forming your initial opinion until that gets modified by new evidence you gather from interactions (likelihood on Data) from data.

   Ex: Choosing an candidate (swing vote)

   if you are an ideologue, there is no likelihood estimation interacting with your past beliefs to change your posterior prediction. There, the likelihood function plays "no role" (i.e no data i.e $\bar{x} = 0$) i.e

   $$\theta_{MAP} \sim \mu \text{ (prior mean)}$$

2. In the absence of prior, MLE is used
3. In the case of flat priors, MAP ~ MLE
4. Use MLE if you have lot of data!
5. In the absence of data, just use prior mean.

Talking points:

(i) MLE of mean of a univariate Gaussian: $\mu_{MLE} = \sum_{i=1}^{n} \frac{x_i}{n}$

MLE of std. dev ---- : $\sigma_{MLE} = \left(\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n} \right)^{1/2}$

(ii) MLE on multiple linear regression (refer to Raj Jain's slides)

$y = Xb + \varepsilon$; b → regression coefficients to determine; ε (Gaussian Noise); X → Data matrix

$$b = (X^T X)^{-1} X^T y$$

"pseudo inverse"

OK, let's look at the general LS problem: $\min \sum_{i=1}^{n} (y_i - a_0 x_i - a_1)^2$

"residuals"

Write the residuals as n-vectors

$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} - \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix}$$

y; X $(n \times 2)$

$\Rightarrow \varepsilon = y - Xb$; in vectorial representation

Therefore, the multiple regression problem i.e LS minimization of the residuals become

$$\|\varepsilon\|^2 = \varepsilon^T \varepsilon = (y - Xb)^T (y - Xb)$$

Objective: Min $(y - Xb)^T (y - Xb)$ wrt "b"

$$\frac{\partial}{\partial b} \{ (y - Xb)^T (y - Xb) \} = X^T y - 2X^T Xb + y^T X$$

$$= \frac{\partial}{\partial b} \{ y^T y - 2b^T X^T y + b^T X^T X b \} = 2X^T y - 2X^T Xb = 2X^T (y - Xb) = 0$$

$$\Rightarrow -2X^T y + 2X^T Xb = 0$$

$$\Rightarrow X^T Xb = X^T y$$

$$\Rightarrow b = (X^T X)^{-1} X^T y$$

MLE for Gaussian noise: $b = (X^T X)^{-1} X^T y$ ← regression coefficients

$\frac{\partial}{\partial b} \{ b^T C b \} = (C + C^T) b$

$C = X^T X \Rightarrow C^T = (X^T X)^T = X^T X = C$

$(y - Xb)^T (y - Xb) = ((Xb)^T + y^T)(y - Xb)$

$= -b^T X^T y + b^T X^T Xb + y^T y - y^T Xb$

$= y^T y - 2b^T X^T y + b^T X^T X b$

$X^T y = (2 \times n)(n \times 1) = 2 \times 1$ ✓

$y^T X = (1 \times n)(n \times 2) = (1 \times 2)$

$X^T y = (y^T X)^T$

Ex:2: Given, samples $\{0,1,0,0,1,0\}$ from a binomial distribution

$p(x=0) = 1-\mu$, $p(x=1) = \mu$; $\mu_{MLE} = ?$

$$L(\mu) = p(x=0)\,p(x=1)\,p(x=0)\,p(x=0)\,p(x=1)\,p(x=0)$$

$$= (1-\mu)^4 \mu^2$$

$$\log L(\mu) = 4 \ln(1-\mu) + 2 \ln \mu$$

$$\frac{1}{\mu}\frac{\partial L}{\partial \mu} = \frac{-4}{1-\mu} + \frac{2}{\mu}$$

$$\Rightarrow \frac{\partial L}{\partial \mu} = \mu\left(-\frac{4}{1-\mu} + \frac{2}{\mu}\right) = 0 \Rightarrow \frac{2}{\mu} = \frac{4}{1-\mu}$$

$$\Rightarrow 1-\mu = 2\mu$$

$$\Rightarrow \boxed{\mu = 1/3}$$

Naive Bayes

Goal: To build a classifier that says whether a text is about sports or not. Training data has 5 sentences:

Data	Label
"A great game"	Sports
"The election was over"	Not sports
"Very clean match"	sports
"A clean but forgettable game"	sports
"It was a close election"	Not sports

Test: A very close game → sports / ¬sports (?)

Notes Naive Bayes is a probabilistic classifier and is a special case of Bayes' Rule. We would like to compute

P(sports | a very close game) & P(¬sports | a very close game)

& pick the one of the larger value.

Step I. Feature Engineering → extract useful features from data for use in the classification problem & remove the ones w/ no bearing on the model (such as identifier information)

Text as data ⇒ "word frequencies" ⇒ ignore word orders & sentence construction, treating every document as a set of the words. So, features ≡ counts of these words.

Step II: Rewrite Bayes Rule in the context of the problem:

$$\underbrace{P(\text{sports} \mid \text{a very close game})}_{P_1} = \frac{P(\text{a very close game} \mid \text{sports}) \times P(\text{sports})}{P(\text{a very close game})}$$

& $$\underbrace{P(\neg\text{sports} \mid \text{a very close game})}_{P_2} = \frac{P(\text{a very close game} \mid \neg\text{sports}) \times P(\neg\text{sports})}{P(\text{a very close game})}$$

i.e discard the denominator as they are same! We need to compute

$p_1 \sim P(\text{a very close game} \mid \text{sports}) \times P(\text{sports})$ &

$p_2 \sim P(\text{a very close game} \mid \neg\text{sports}) \times P(\neg\text{sports})$

NOTE: "a very close game" doesn't appear in the training data.

↳ Naive handling of the above problem ←

$$P(\text{a very close game}) = P(a) \times P(\text{very}) \times P(\text{close}) \times P(\text{game})$$

strong assumption

Therefore,

(a) $\Rightarrow P(\text{a very close game} \mid \text{sports}) = \cancel{P(a) \times P}\ P(a \mid \text{sports}) \times P(\text{very} \mid \text{sports}) * P(\text{close} \mid \text{sports}) * P(\text{game} \mid \text{sports})$

Compute these

Similarly,

(b) $\Rightarrow P(\text{a very close game} \mid \neg\text{sports}) = P(a \mid \neg\text{sports}) * P(\text{very} \mid \neg\text{sports}) * P(\text{close} \mid \neg\text{sports}) * P(\text{game} \mid \neg\text{sports})$

Compute these

Step III: $P(\text{sports}) = 3/5$, $P(\neg\text{sports}) = \frac{2}{5}$; apriori

$$P(\text{game} \mid \text{sports}) = \frac{\text{word frequency of 'game' in sports}}{\text{total \# of words in sports}} = \frac{2}{11}$$

$P(\text{close} \mid \text{sports}) = 0$! → how do we handle this ⇒ add 1 to every count so that it's never zero, → to balance this we add the # of possible words to the divisor, so that the ratio ≤ 1. In this case, # of possible words = 14

$P(a \mid sports) = \frac{2+1}{11+14}$; $P(a \mid \neg sports) = \frac{1+1}{9+14}$

$P(very \mid sports) = \frac{1+1}{11+14}$; $P(very \mid \neg sports) = \frac{0+1}{11+14}$

$P(close \mid sports) = \frac{0+1}{11+14}$; $P(close \mid \neg sports) = \frac{1+1}{9+14}$

$P(game \mid sports) = \frac{2+1}{11+14}$; $P(game \mid \neg sports) = \frac{0+1}{9+14}$

$\therefore p_1 = 0.0000276$; $p_2 = 0.00000572$

$\searrow$ $p(sports \mid a\ very\ close\ game) > p(\neg sports \mid a\ very\ close\ game)$

$\Downarrow$

Test data label : sports