Bayes Classification

Alymzhan Toleu
alymzhan.toleu@gmail.com

Axioms of probability (Kolmogorov's axioms)

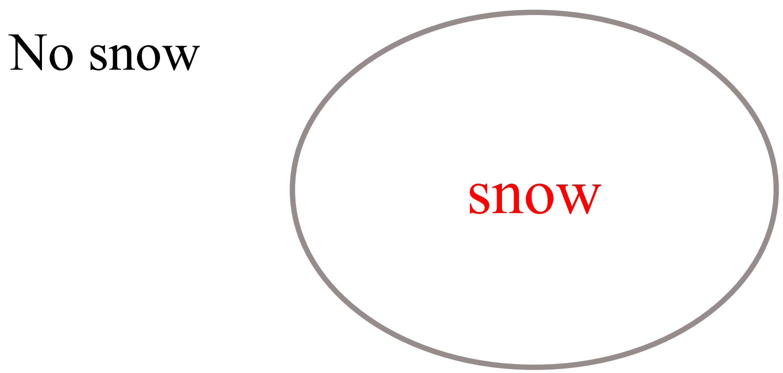
A variety of useful facts can be derived from just three axioms:

- $0 \le P(A) \le 1$
- P(true) = 1, P(false) = 0
- Union: The union of two events is the probability that either A or B will occur.
  - $P(A \cup B) = P(A) + P(B) - P(A \cup B)$
- Intersection: The intersection of two events is the probability that the two events, A and B, will occur at the same time.
  - Independent events: $P(A \cap B) = P(A) * P(B)$

Two events are mutually exclusive if they cannot occur at the same time.

Prior

- The **prior probability** of an event refers to the degree of belief assigned to that event before incorporating any additional information.

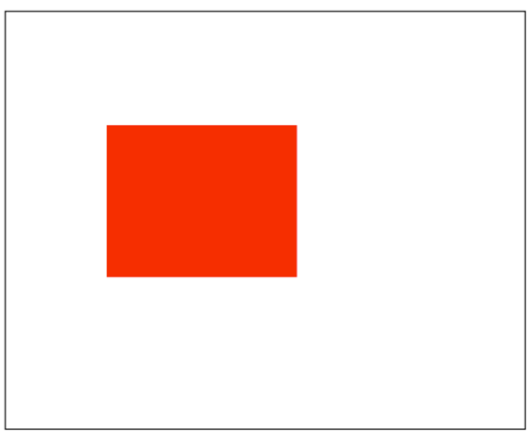


P(snow tomorrow) = 0.2
P(no snow tomorrow) = 0.8

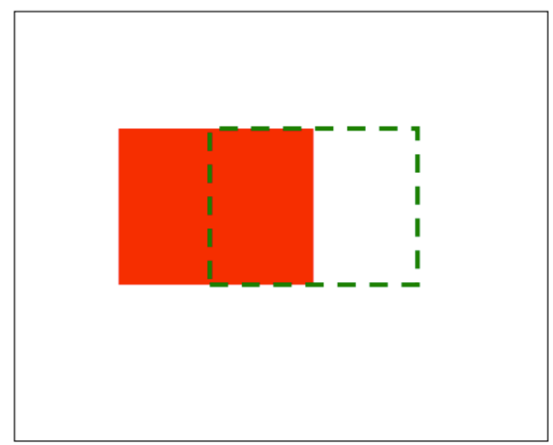
Conditional probability

- P(A = true| B = false): The fraction of cases where A is true if B is false.
- For example:

P(A = 0.2)



P(A|B = 0.5)



Conditional probability

- The prior belief of a random variable can be improved in some cases by conditioning on one or more other random variables.
- For example:

$$\text{P(slept in moive)} = \frac{4}{7}$$

$$\text{P(slept in moive | liked movie)} = \frac{1}{4}$$

$$\text{P(did not slept in moive | liked movie)} = \frac{3}{4}$$

Slept	Liked
1	1
0	1
1	0
1	0
0	1
1	0
0	1

Joint distributions

- The probability that a set of random variables will take a specific value is their joint distribution.
- Notation: $P(A \cap B)$ or $P(A,B)$
- Example: P(liked movie, slept) = ?

If we assume independence then

$$P(A,B)=P(A)P(B)$$

However, in many cases such an assumption maybe too strong

Joint distributions

$$\mathrm{P}(\text{Length} = \text{short}) = \frac{3}{7}$$

$$\mathrm{P}(\text{slept in moive}) = \frac{4}{7}$$

P(Length = short , slept in moive) = ?

Length	**Slept**	**Liked**
Short	1	1
Long	0	1
Medium	1	0
Short	1	0
Medium	0	1
Short	1	0
Long	0	1

Joint distributions

$$\mathrm{P(Length = short)} = \frac{3}{7} = 0.42$$

$$\mathrm{P(slept\ in\ moive)} = \frac{4}{7} = 0.57$$

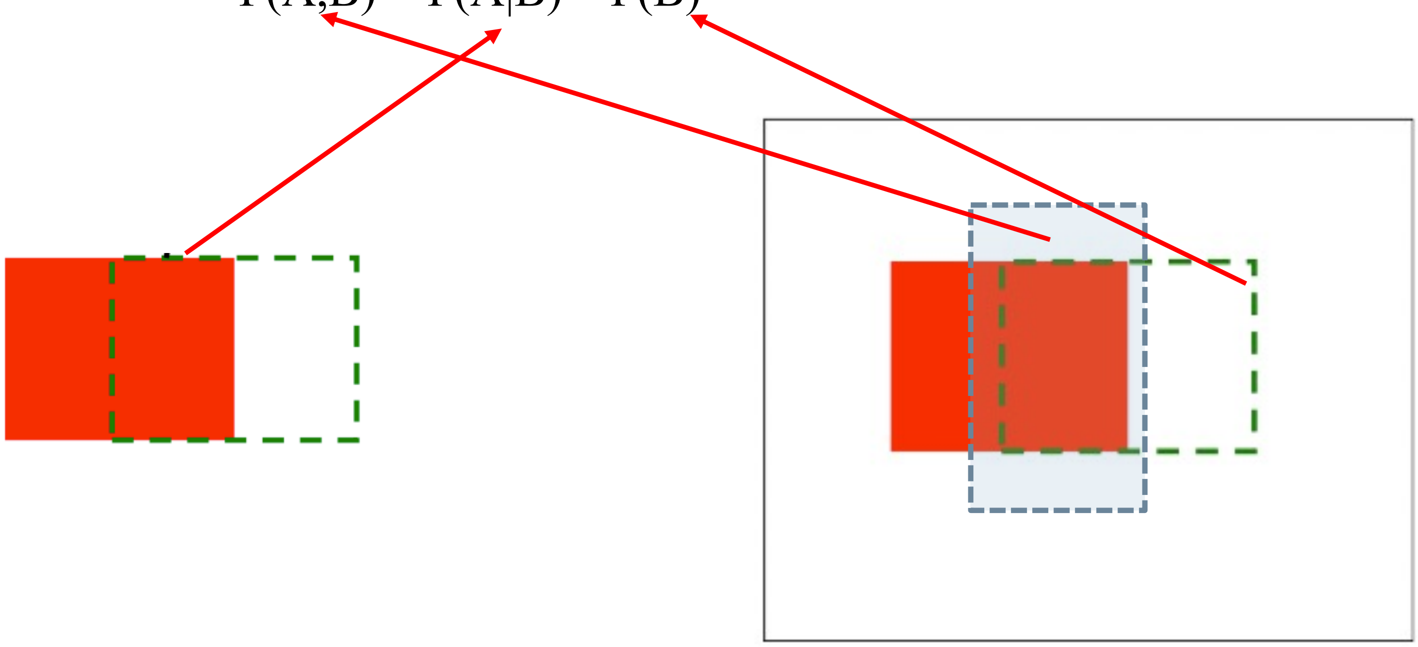
$$\mathrm{P(Length = short\ ,\ slept\ in\ moive)} = \frac{3}{7} = 0.42$$

Length	Slept	Liked
Short	1	1
Long	0	1
Medium	1	0
Short	1	0
Medium	0	1
Short	1	0
Long	0	1

Chain rule

- The joint distribution can be specified in terms of conditional probability:

$$P(A,B) = P(A|B) * P(B)$$



Bayes decision rule

- One of the most important rules in probabilistic theory.
- Derived from the chain rule:

$$P(A,B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

- Thus, it becomes generative models

Joint probabilities

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

Thomas Bayes was an English clergyman who set out his theory of probability in 1764

Bayes decision rule

- If we know the conditional probability P(X | y) we can determine the appropriate class by using Bayes rule:

Posterior

class-conditional probability
Or Likelihood

Prior

$$P(y|X) = \frac{p(X,y)}{p(X)} = \frac{p(X|y) * p(y)}{P(X)}$$

Evidence

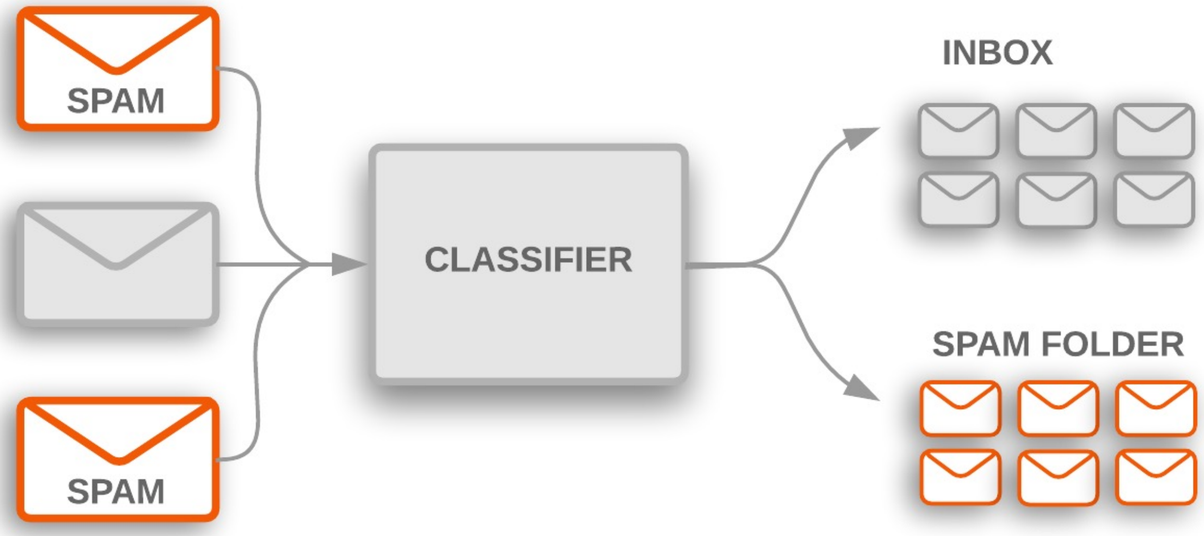
Recall…
X – input attributes (features)
y – the class label

But how do we determine p(X|y)?

Types of classifier

- We can divide the large variety of classification approaches into roughly three main types
  - Discriminative
    - directly estimate a decision rule/boundary - e.g., decision tree
  - Instance based classifiers
    - use observation directly (no models) - e.g. K nearest neighbors
  - Generative
    - build a generative statistical model - e.g., Bayesian networks

Spam detection: Example



Normal Emails (8)

Words	Frequncy
Dear	8
Friend	5
Lunch	3
Money	1

The probability of each word we see in the normal messages:

$$P(\boldsymbol{Dear}|Normal) = \frac{8}{17} = 0.47$$

$$P(\boldsymbol{Friend}|Normal) = \frac{5}{17} = 0.29$$

$$P(\boldsymbol{Lunch}|Normal) = \frac{3}{17} = 0.18$$

$$P(\boldsymbol{Money}|Normal) = \frac{1}{17} = 0.06$$

Normal Emails (8)

Words	Frequncy
Dear	8
Friend	5
Lunch	3
Money	1

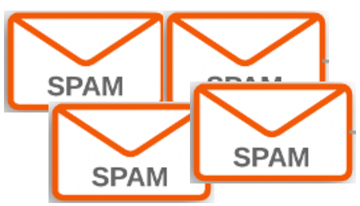
The probability of each word we see in the normal messages:

$$P(\boldsymbol{Dear}|Normal) = \frac{8}{17} = 0.47$$

$$P(\boldsymbol{Friend}|Normal) = \frac{5}{17} = 0.29$$

$$P(\boldsymbol{Lunch}|Normal) = \frac{3}{17} = 0.18$$

$$P(\boldsymbol{Money}|Normal) = \frac{1}{17} = 0.06$$



Spams(4)

Words	Frequncy
Dear	2
Friend	1
Lunch	0
Money	4

Likewaise, we calculate the probability of each word we see in the spam:

$$P(\boldsymbol{Dear}|Spam) = \frac{2}{7} = 0.29$$

$$P(\boldsymbol{Friend}|Spam) = \frac{1}{7} = 0.06$$

$$P(\boldsymbol{Lunch}|Spam) = \frac{0}{7} = 0$$

$$P(\boldsymbol{Money}|Spam) = \frac{4}{7} =0.57$$

Normal Emails (8)

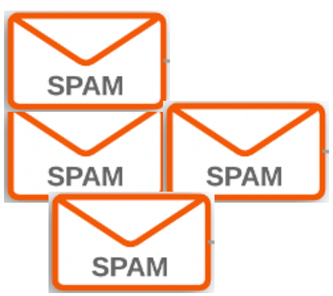
$$P(\boldsymbol{Dear}|Normal) = \frac{8}{17} = 0.47$$

$$P(\boldsymbol{Friend}|Normal) = \frac{5}{17} = 0.29$$

$$P(\boldsymbol{Lunch}|Normal) = \frac{3}{17} = 0.18$$

$$P(\boldsymbol{Money}|Normal) = \frac{1}{17} = 0.06$$

These probabilities can be called **Likelihood.** Because they are probabilties of discrete, and dot probability of something continuous.



Spams (4)

$$P(\boldsymbol{Dear}|Spam) = \frac{2}{7} = 0.29$$

$$P(\boldsymbol{Friend}|Spam) = \frac{1}{7} = 0.06$$

$$P(\boldsymbol{Lunch}|Spam) = \frac{0}{7} = 0$$

$$P(\boldsymbol{Money}|Spam) = \frac{4}{7} = 0.57$$

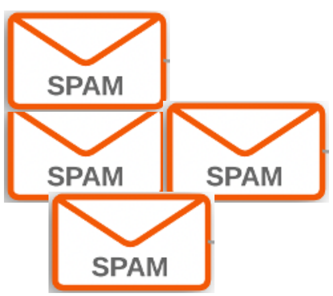
Normal Emails

$$P(\boldsymbol{Dear}|Normal) = \frac{8}{17} = 0.47$$

$$P(\boldsymbol{Friend}|Normal) = \frac{5}{17} = 0.29$$

$$P(\boldsymbol{Lunch}|Normal) = \frac{3}{17} = 0.18$$

$$P(\boldsymbol{Money}|Normal) = \frac{1}{17} = 0.06$$



Spams

$$P(\boldsymbol{Dear}|Spam) = \frac{2}{7} = 0.29$$

$$P(\boldsymbol{Friend}|Spam) = \frac{1}{7} = 0.06$$

$$P(\boldsymbol{Lunch}|Spam) = \frac{0}{7} = 0$$

$$P(\boldsymbol{Money}|Spam) = \frac{4}{7} =0.57$$

Let’s say, we got a new message that said:

Dear Friend

We want to classify if it is a Normal or Spam message.

What we do?

$P(N) * P(\boldsymbol{Dear}|N) * P(\boldsymbol{Friend}|N) = 0.09$

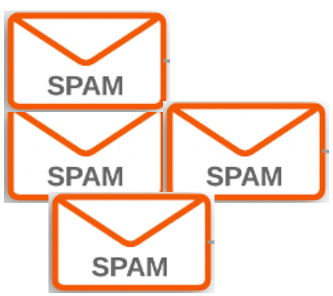
Normal Emails (8)

$$P(\boldsymbol{Dear}|N) = \frac{8}{17} = 0.47$$

$$P(\boldsymbol{Friend}|N) = \frac{5}{17} = 0.29$$

$$P(\boldsymbol{Lunch}|N) = \frac{3}{17} = 0.18$$

$$P(\boldsymbol{Money}|N) = \frac{1}{17} = 0.06$$



Spams (4)

$$P(\boldsymbol{Dear}|S) = \frac{2}{7} = 0.29$$

$$P(\boldsymbol{Friend}|S) = \frac{1}{7} = 0.06$$

$$P(\boldsymbol{Lunch}|S) = \frac{0}{7} = 0$$

$$P(\boldsymbol{Money}|S) = \frac{4}{7} = 0.57$$

Let’s say, we got a new message that said:

Dear Friend ………… ………

First, we need to calculate the initial guess (**prior probability**) of a message, regardless what it says, is a normal message.

$$P(N) = \frac{8}{8+4} = 0.66$$

$$P(N) * P(\boldsymbol{Dear}|N) * P(\boldsymbol{Friend}|N)$$
$$= 0.66. * 0.47 * 0.29 = \boxed{0.09}$$

It is the score is given to the message “Dear Friend” if it is a normal.

$P(N) * P(\boldsymbol{Dear}|N) * P(\boldsymbol{Friend}|N) = 0.09$

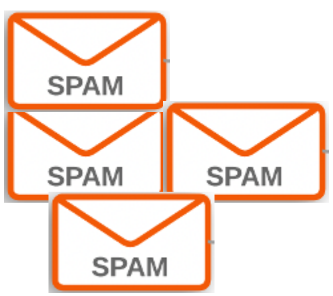
Normal Emails (8)

$$P(\boldsymbol{Dear}|N) = \frac{8}{17} = 0.47$$

$$P(\boldsymbol{Friend}|N) = \frac{5}{17} = 0.29$$

$$P(\boldsymbol{Lunch}|N) = \frac{3}{17} = 0.18$$

$$P(\boldsymbol{Money}|N) = \frac{1}{17} = 0.06$$



Spams (4)

$$P(\boldsymbol{Dear}|S) = \frac{2}{7} = 0.29$$

$$P(\boldsymbol{Friend}|S) = \frac{1}{7} = 0.06$$

$$P(\boldsymbol{Lunch}|S) = \frac{0}{7} = 0$$

$$P(\boldsymbol{Money}|S) = \frac{4}{7} = 0.57$$

Let’s say, we got a new message that said:

> Dear Friend
> ………..
> ……..

First, we need to calculate the initial guess **(prior probability)** of a message, regardless what it says, is a Spam message.

$$P(S) = \frac{4}{8+4} = 0.33$$

$$P(S) * P(\boldsymbol{Dear}|S) * P(\boldsymbol{Friend}|S)$$
$$= 0.33 * 0.29 * 0.06 = \boxed{0.01}$$

It is the score is given to the message “Dear Friend” if it is a Spam.

$P(N) * P(\boldsymbol{Dear}|N) * P(\boldsymbol{Friend}|N) = 0.09$

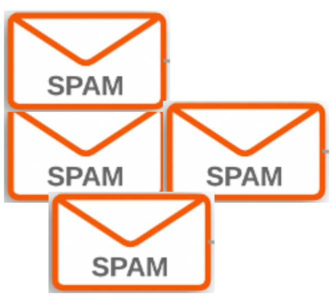
Normal Emails (8)

$$P(\boldsymbol{Dear}|N) = \frac{8}{17} = 0.47$$

$$P(\boldsymbol{Friend}|N) = \frac{5}{17} = 0.29$$

$$P(\boldsymbol{Lunch}|N) = \frac{3}{17} = 0.18$$

$$P(\boldsymbol{Money}|N) = \frac{1}{17} = 0.06$$



Spams (4)

$$P(\boldsymbol{Dear}|S) = \frac{2}{7} = 0.29$$

$$P(\boldsymbol{Friend}|S) = \frac{1}{7} = 0.06$$

$$P(\boldsymbol{Lunch}|S) = \frac{0}{7} = 0$$

$$P(\boldsymbol{Money}|S) = \frac{4}{7} = 0.57$$

Let’s say, we got a new message that said:

Dear Friend
………..
……..

$$P(N) * P(\boldsymbol{Dear}|N) * P(\boldsymbol{Friend}|N) = 0.09$$

$$P(S) * P(\boldsymbol{Dear}|S) * P(\boldsymbol{Friend}|S) = 0.01$$

$0.09\ (N) > 0.01\ (S)$

This is a normal message.

This is the basics how **Naïve Bayes Classification** works.

The problem of zero probability

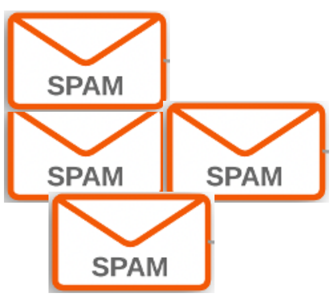
Normal Emails (8)

$$P(\boldsymbol{Dear}|N) = \frac{8}{17} = 0.47$$

$$P(\boldsymbol{Friend}|N) = \frac{5}{17} = 0.29$$

$$P(\boldsymbol{Lunch}|N) = \frac{3}{17} = 0.18$$

$$P(\boldsymbol{Money}|N) = \frac{1}{17} = 0.06$$



Spams (4)

$$P(\boldsymbol{Dear}|S) = \frac{2}{7} = 0.29$$

$$P(\boldsymbol{Friend}|S) = \frac{1}{7} = 0.06$$

$$P(\boldsymbol{Lunch}|S) = \frac{0}{7} = 0$$

$$P(\boldsymbol{Money}|S) = \frac{4}{7} = 0.57$$

Let's say, we got a new message that said:

> Lunch Money Money Money Money
> ………..

$$P(N) * P(\boldsymbol{Lunch}|N) * P(\boldsymbol{Money}|N)^4 = 0.000002$$

$$P(S) * P(\boldsymbol{Lunch}|S) * P(\boldsymbol{Money}|S)^4 = 0$$

$$0.00002(N) > 0\ (\text{S})$$

This is a normal message.

The problem of zero probability

- To avoid this issue, there is an approach called **smoothing technique.**
- A small number of counts (α, alpha) will add to each sample (word).
- Make sure there are no zero probability.

Smoothing technique

$set\ \alpha = 1$

Normal Emails (8)

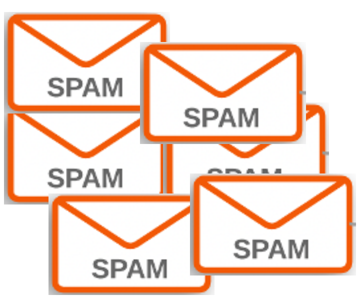
Words	Frequncy
Dear	$8+\alpha$
Friend	$5+\alpha$
Lunch	$3+\alpha$
Money	$1+\alpha$

$$P(\boldsymbol{Dear}|N) = \frac{9}{21} = 0.43$$

$$P(\boldsymbol{Friend}|N) = \frac{6}{21} = 0.29$$

$$P(\boldsymbol{Lunch}|N) = \frac{4}{21} = 0.19$$

$$P(\boldsymbol{Money}|N) = \frac{2}{21} = 0.1$$



Spams(4)

Words	Frequncy
Dear	$2+\alpha$
Friend	$1+\alpha$
Lunch	$0+\alpha$
Money	$4+\alpha$

$$P(\boldsymbol{Dear}|S) = \frac{3}{11} = 0.27$$

$$P(\boldsymbol{Friend}|S) = \frac{2}{11} = 0.18$$

$$P(\boldsymbol{Lunch}|S) = \frac{1}{11} = 0.09$$

$$P(\boldsymbol{Money}|S) = \frac{5}{11} = 0.45$$

Smoothing technique

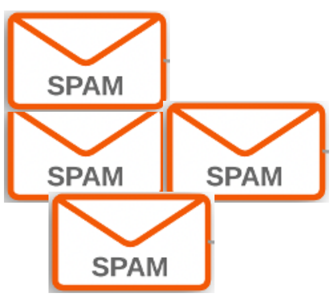
Normal Emails (8)

$$P(\boldsymbol{Dear}|N) = \frac{9}{21} = 0.43$$

$$P(\boldsymbol{Friend}|N) = \frac{6}{21} = 0.29$$

$$P(\boldsymbol{Lunch}|N) = \frac{4}{21} = 0.19$$

$$P(\boldsymbol{Money}|N) = \frac{2}{21} = 0.1$$



Spams (4)

$$P(\boldsymbol{Dear}|S) = \frac{3}{11} = 0.27$$

$$P(\boldsymbol{Friend}|S) = \frac{2}{11} = 0.18$$

$$P(\boldsymbol{Lunch}|S) = \frac{1}{11} = 0.09$$

$$P(\boldsymbol{Money}|S) = \frac{5}{11} = 0.45$$

Let's say, we got a new message that said:

Lunch Money Money Money Money
………..

$$P(N) * P(\boldsymbol{Lunch}|N) * P(\boldsymbol{Money}|N)^4 = 0.00001$$

$$P(S) * P(\boldsymbol{Lunch}|S) * P(\boldsymbol{Money}|S)^4 = 0.00122$$

$$0.00001(N) < 0.00122\ (\mathrm{S})$$

This is a spam message.

- Why Naïve Bayes Classification is **Naïve?**

Word order issue

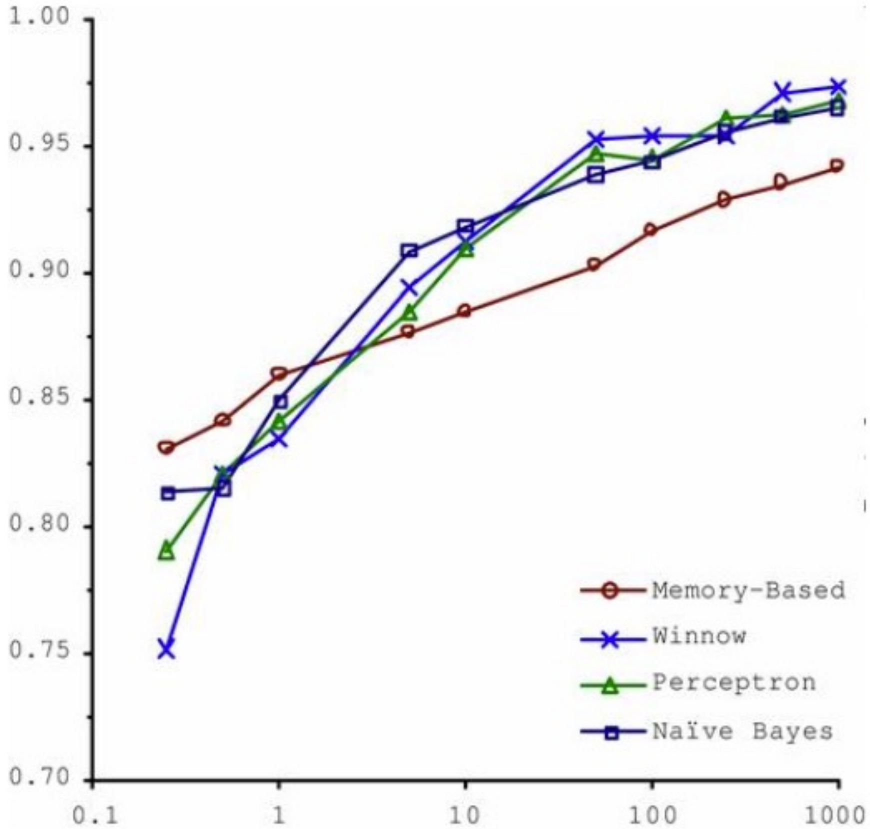
- It treats each email as a bag of words.
- In other words, it treats all word orders the same.
- It ignores word orders.
- For example:
  - Score fir **Dear Friend** and **Friend Dear** are same.

$P(N) * P(\boldsymbol{Dear}|N) * P(\boldsymbol{Friend}|N)$ =0.09

$P(N) * P(\boldsymbol{Friend}|N) * P(\boldsymbol{Dear}|N)$ =0.09

Data size is matter

Accuracy



Data size

Maximum Likelihood Estimation

Likelihood vs. Probability

- **Probability** corresponds to finding the chance of something given a sample distribution of the data.

- **Likelihood** refers to finding the best distribution of the data given a particular value of some feature or some situation in the data.

Probability $P(data \mid distribution)$

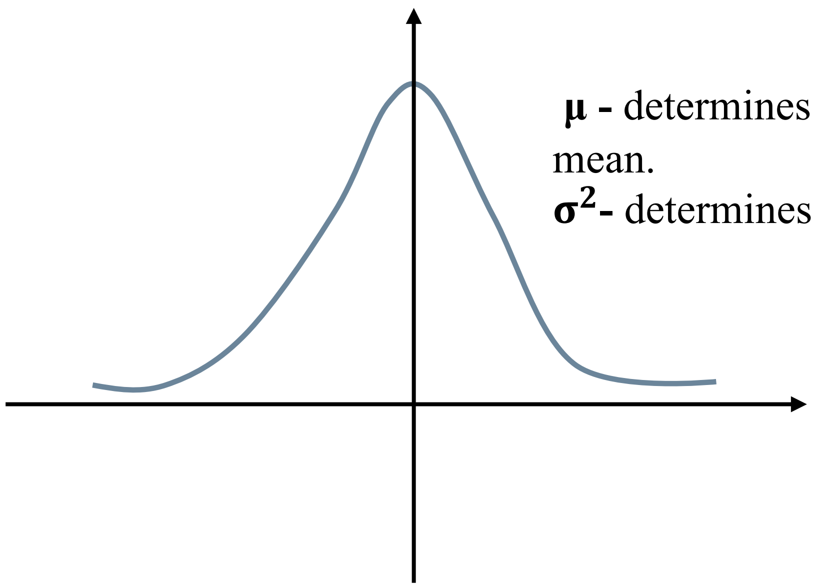
Likelihood $\text{Likelihood}(distribution \mid data)$

The Normal Distribution

- The Gaussian, also known as the normal distribution, is a widely used model for the distribution of continuous variables.

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

where μ is the mean and σ^2 is the variance.



μ - determines the location of the normal distribution's mean.
$\boldsymbol{\sigma^2}$**-** determines the normal distribution's width.

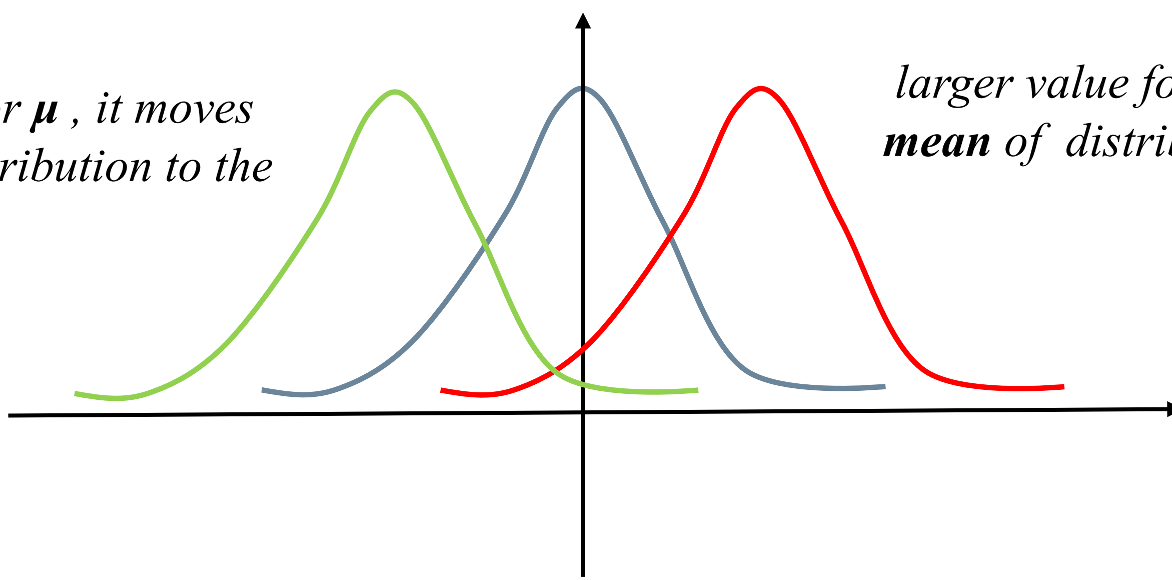
The Normal Distribution

- The Gaussian, also known as the normal distribution, is a widely used model for the distribution of continuous variables.

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

where μ is the mean and σ^2 is the variance.

Smaller value for ***μ*** *, it moves the* ***mean*** *of distribution to the left*



larger value for ***μ*** *, it moves the* ***mean*** *of distribution to the right*

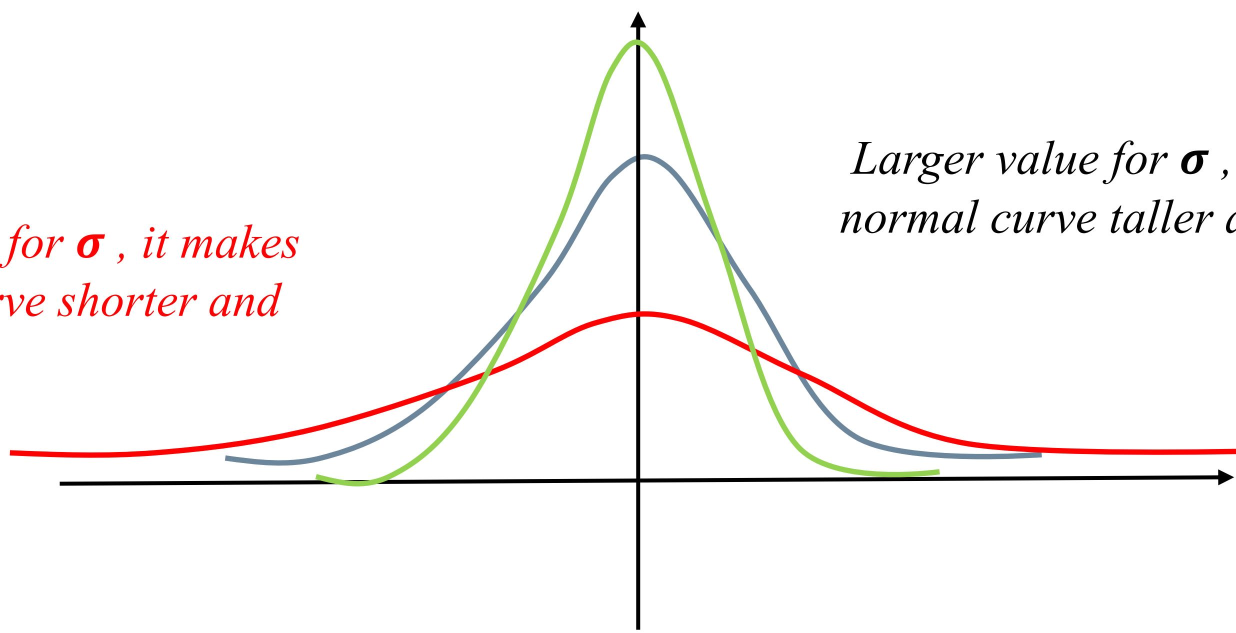
The Normal Distribution

- The Gaussian, also known as the normal distribution, is a widely used model for the distribution of continuous variables.

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

where μ is the mean and σ^2 is the variance.

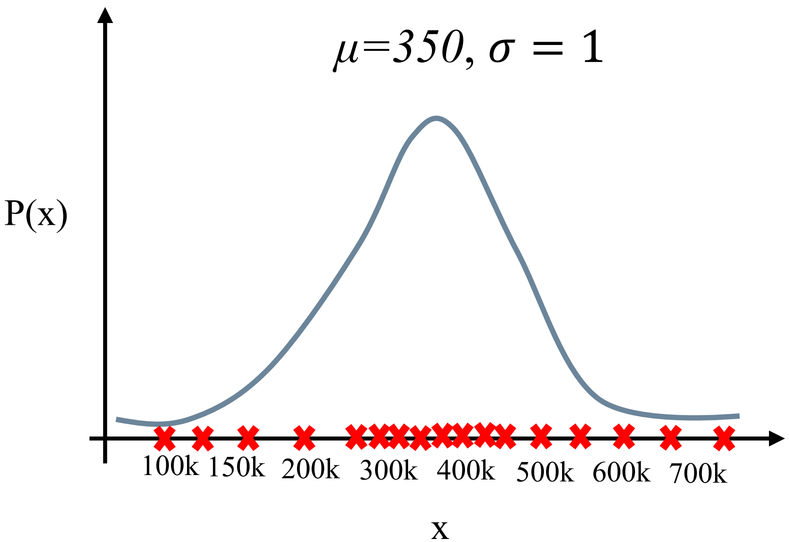
Smaller value for $\boldsymbol{\sigma}$ *, it makes the normal curve shorter and wider.*



Larger value for $\boldsymbol{\sigma}$ *, it makes the normal curve taller and narrower.*

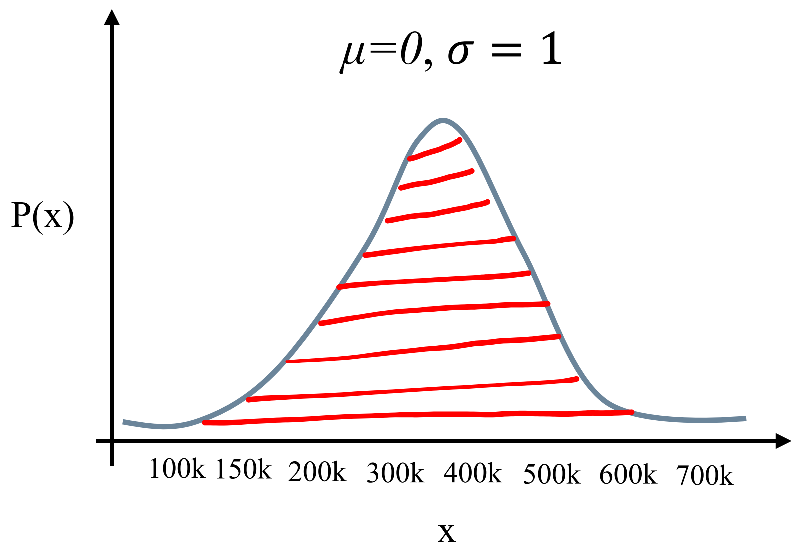
Likelihood vs. Probability: Example

Distribution of employee's income in a company.



Likelihood vs. Probability: Example

Distribution of employee's income in a company.

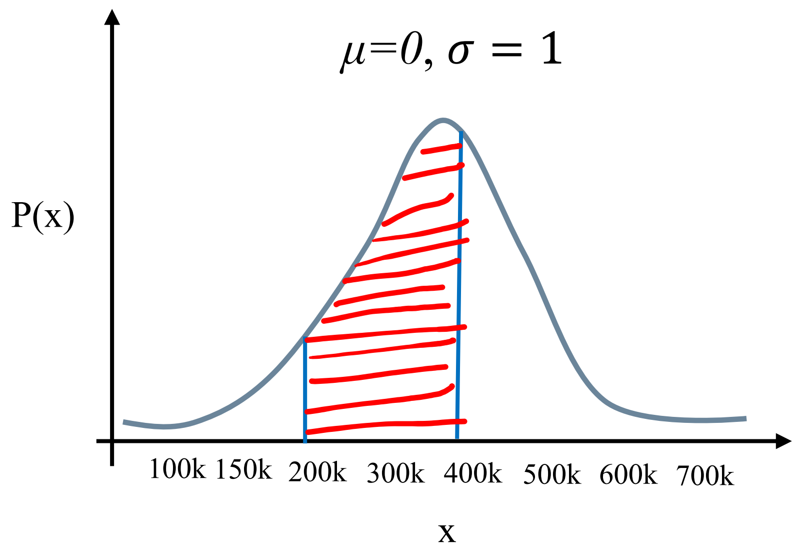


$$\int p(x) = 1$$

Total area under the curve.

Likelihood vs. Probability: Example

Distribution of employee's income in a company.



Probability of employee's income between 200k and 400k:

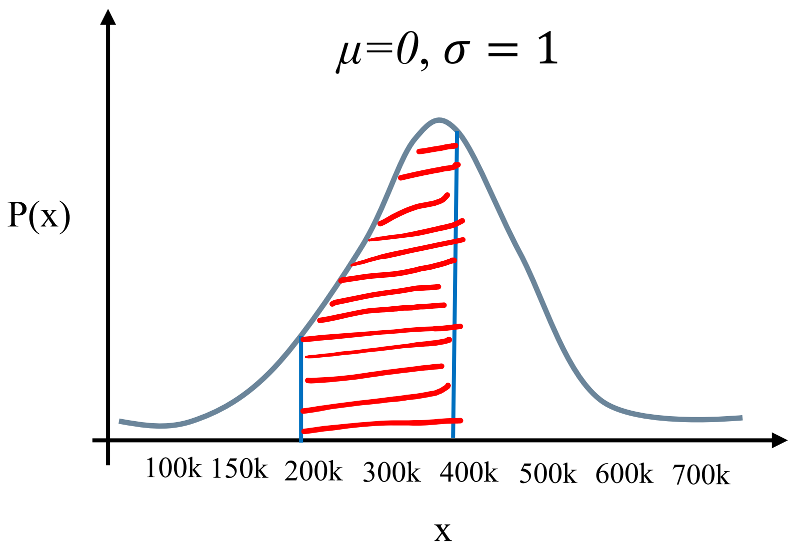
$$\int_{200k}^{400k} p(x)dx$$

What is the relation between likelihood and probability here?

Likelihood vs. Probability: Example

Probability

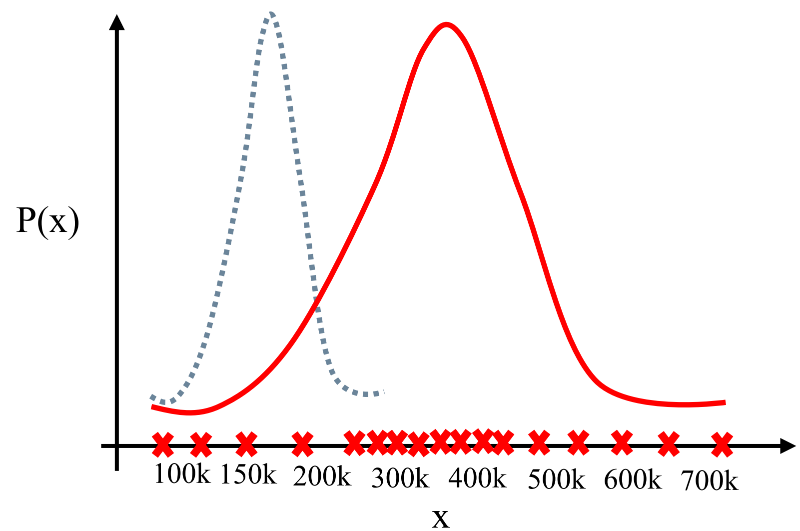
Distribution of income in an organization.



Probability of employee's income between 200k and 400k:

$$\int_{200k}^{400k} p(x)dx$$

Likelihood



We want to find some distribution fits to this data.

μ=?, σ =?

For example:

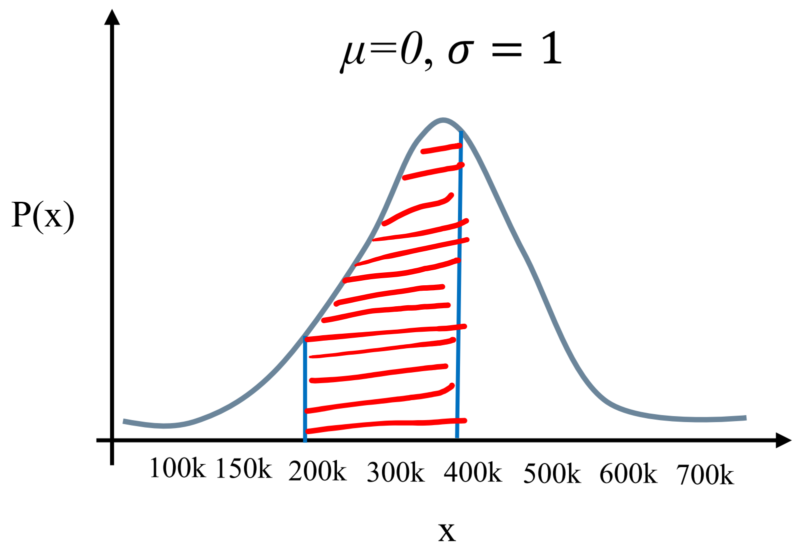
μ=10, $\sigma = 0.5$

μ=20, $\sigma = 2$

Likelihood vs. Probability: Example

Probability

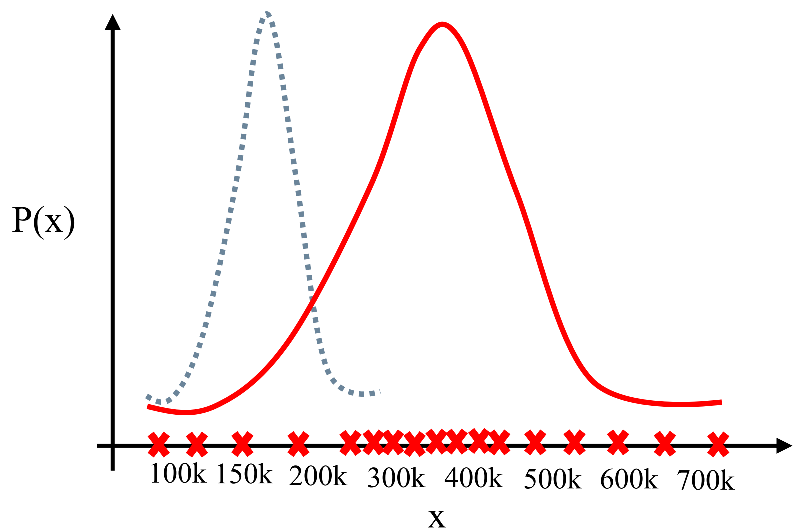
Distribution of income in an organization.



The probability of employee's income between 200k and 400k:

$$\int_{200k}^{400k} p(x)dx$$

Likelihood



We want to find some distribution fit to this data.

μ=?, σ =?

Add likelihood function:

$\boldsymbol{L}$ (μ=10, $\sigma = 0.5$) = ?

$\boldsymbol{L}$ (μ=20, $\sigma = 2$) = ?

Definition: Likelihood

"The likelihood that any parameter (or set of parameters) should have any assigned value (or set of values) is proportional to the probability that if this were so, the totality of observations should be that observed."

— Fisher, 1922

Ronald Fisher

Definition: Likelihood

"The likelihood that any parameter (or set of parameters) should have any assigned value (or set of values) is proportional to the probability that if this were so, the totality of observations should be that observed."

— Fisher, 1922

Ronald Fisher

$$\boldsymbol{L}(\mu{=}10, \sigma = 0.5) \propto P\left(x^{(1)} = 100k, x^{(2)} = 200k, x^{(3)} = 300k \ldots x^{(n)} \middle| \mu{=}10, \sigma = 0.5\right)$$

We pass in the parameters of the distribution

The likelihood value going to be proportional to the probability of observing all of these examples given the parameters of the assumed distribution

Maximize the Likelihood

- Find the values of theta that is going to maximize the likelihood function.

$$\hat{\theta}^{MLE} = argmax\ \boldsymbol{L}(\theta)$$

Maximize the Likelihood

- Likelihood function

$$\boldsymbol{L}(\theta) \propto P(D \mid \theta) \quad \Rightarrow \quad \hat{\theta}^{MLE} = argmax\, \boldsymbol{L}(\theta)$$

$$\boldsymbol{L}(\theta) \propto P\left(x^{(1)} = 100k, x^{(2)} = 200k, x^{(3)} = 300k \ldots x^{(n)} \middle| \theta\right)$$

Assume these incomes are independent and identically distributed (i.i.d.).

$$\boldsymbol{L}(\theta) \propto P\left(x^{(1)} \middle| \theta\right) * P\left(x^{(2)} \middle| \theta\right) * P\left(x^{(3)} \middle| \theta\right) * \cdots * P\left(x^{(n)} \middle| \theta\right))$$

$$\boldsymbol{L}(\theta) \propto \prod_{\boldsymbol{1}}^{\boldsymbol{n}} P\left(x^{(i)} \middle| \theta\right)$$

Multiplications lead to the arithmetic underflow

The term arithmetic underflow is a condition in a computer program where the result of a calculation is a number of more precise absolute value than the computer can actually represent in memory on its central processing unit (CPU).

Maximize the Likelihood

- Likelihood function

$$\boldsymbol{L}(\theta) \propto P(D \mid \theta) \quad \Rightarrow \quad \hat{\theta}^{MLE} = argmax\, \boldsymbol{L}(\theta)$$

$$\boldsymbol{L}(\theta) \propto P\left(x^{(1)} = 100k, x^{(2)} = 200k, x^{(3)} = 300k \ldots x^{(n)} \middle|\, \theta\right)$$

Assume these incomes are independent and identically distributed (i.i.d.).

$$\boldsymbol{L}(\theta) \propto P\left(x^{(1)} \middle|\, \theta\right) * P\left(x^{(2)} \middle|\, \theta\right) * P\left(x^{(3)} \middle|\, \theta\right) * \cdots * P\left(x^{(n)} \middle|\, \theta\right))$$

$$\boldsymbol{L}(\theta) \propto \prod_{\boldsymbol{1}}^{\boldsymbol{n}} P\left(x^{(i)} \middle|\, \theta\right)$$

Multiplications lead to the arithmetic underflow

Taking the logarithms on both sides: $\boldsymbol{log}(\boldsymbol{L}(\theta)) \propto \boldsymbol{log}(\prod_{\boldsymbol{i=1}}^{\boldsymbol{n}} P\left(x^{(i)} \middle|\, \theta\right))$

$$\boldsymbol{log}(\boldsymbol{L}(\theta)) \propto \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{log}(P\left(x^{(i)} \middle|\, \theta\right))$$

Maximize the Likelihood

- Likelihood function

$$\boldsymbol{L}(\theta) \propto P(D \mid \theta) \quad \Rightarrow \quad \hat{\theta}^{MLE} = argmax\, \boldsymbol{L}(\theta)$$

$$\Downarrow \quad \boldsymbol{log}(\boldsymbol{L}(\theta)) \propto \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{log}\left(P\left(x^{(i)} \middle| \, \theta\right)\right)$$

$$\hat{\theta}^{MLE} = \operatorname{argmax} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{log}\left(P\left(x^{(i)} \middle| \, \theta\right)\right)$$

Maximize Likelihood Estimation

Maximum Likelihood
For the Normal Distribution

Maximum Likelihood For the Normal Distribution

- the Normal Distribution

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

where μ is the mean and σ^2 is the variance.

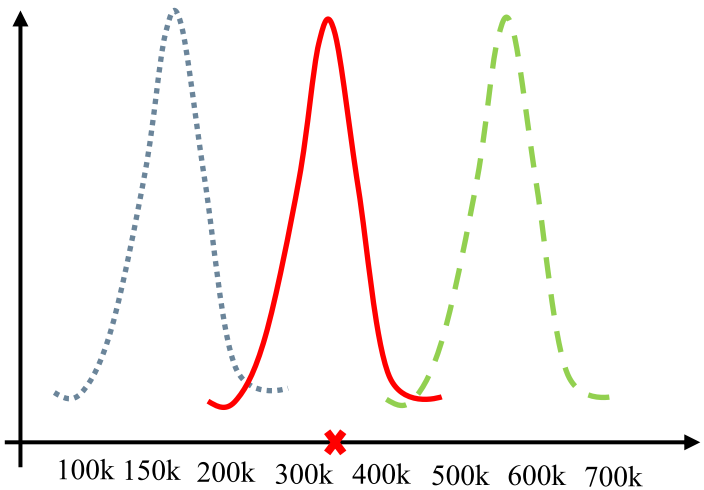
- the Likelihood of the normal Distribution

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

Example (cont.)

Distribution of income in an organization.

For simplicity, we assume there is only one employee: $x^{(1)}=350$



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

$L\ (\mu=100,\ \sigma = 0.5) = ?$

$L\ (\mu=300,\ \sigma = 0.5) = ?$

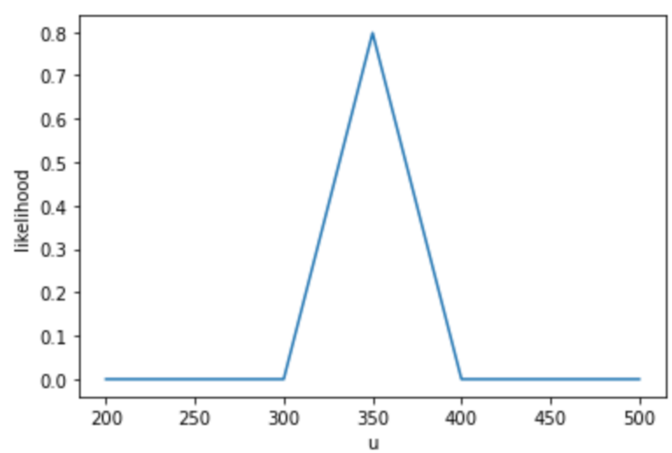
$L\ (\mu=500,\ \sigma = 0.5) = ?$

```python
mus = [200,300,350,400,500]
sigma = 0.5

pvs = []
for mu in mus:
    pv = scipy.stats.norm(mu, sigma).pdf(35
    pvs.append(pv)
```

```python
plt.xlabel("u")
plt.ylabel("likelihood")
plt.plot(mus,pvs)
```

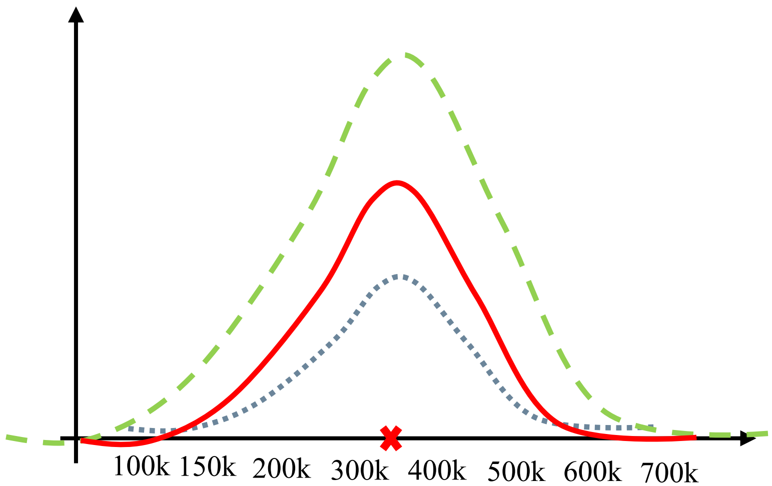
```
[<matplotlib.lines.Line2D at 0x166b06490>]
```



Example (cont.)

Distribution of income in an organization.

For simplicity, we assume there is only one employee: $x^{(1)}=350$



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

$L\ (\mu=200,\ \sigma = 0.5) = ?$

$L\ (\mu=200,\ \sigma = 1) = ?$

$L\ (\mu=200,\ \sigma = 2) = ?$

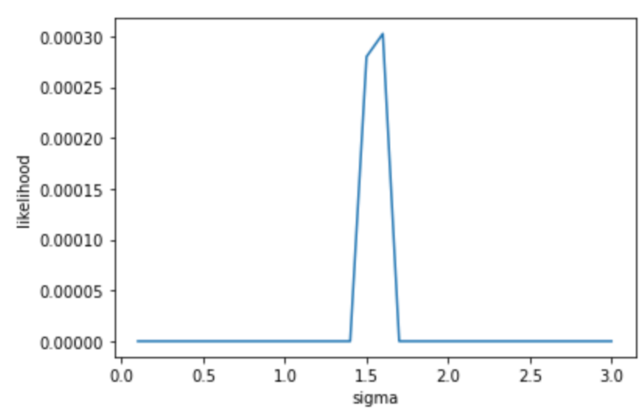
```
]:
sigmas = np.linspace(start=0.1, stop=3, num=30)

xvs = np.linspace(start=100, stop=700, num=30,dtype=int)
mu = np.sum(xvs)/len(xvs)

pvs = []
for x in xvs:
    _pvs = []
    for sigma in sigmas:
        pv = scipy.stats.norm(mu, sigma).pdf(x)
        _pvs.append(pv)
    pvs.append(np.max(_pvs))
```

```
]:
plt.xlabel("sigma")
plt.ylabel("likelihood")
plt.plot(sigmas,pvs)
```

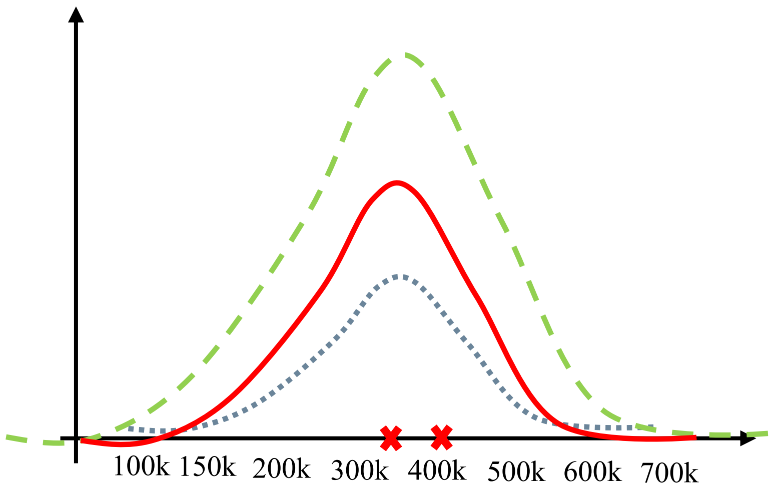
```
]: [<matplotlib.lines.Line2D at 0x166b38bb0>]
```



Example (cont.)

Distribution of income in an organization.

For simplicity, we assume there are two employees: : $x^{(1)}$=350 and $x^{(2)}$=400



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

$$L(\mu, \sigma \mid x^{(1)} = 350, x^{(2)} = 400)$$

$$= L(\mu, \sigma \mid x^{(1)} = 350) \ * L(\mu, \sigma \mid x^{(2)} = 400)$$

If there are *n* employees

$$L(\mu, \sigma \mid x^{(1)}, x^{(2)}, \dots, x^{(n)}) \ = L(\mu, \sigma \mid x^{(1)}) \ * L(\mu, \sigma \mid x^{(2)}) * \cdots * L(\mu, \sigma \mid x^{(n)})$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x^{(1)}-\mu)^2}{2\sigma^2}} * \cdots * \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x^{(n)}-\mu)^2}{2\sigma^2}}$$

Example (cont.)

$$L(\mu, \sigma \mid x^{(1)}, x^{(2)}, \dots, x^{(n)})$$

$$= L(\mu, \sigma \mid x^{(1)}) \ * \ L(\mu, \sigma \mid x^{(2)}) * \cdots * L(\mu, \sigma \mid x^{(n)})$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x^{(1)}-\mu)^2}{2\sigma^2}} * \cdots * \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x^{(n)}-\mu)^2}{2\sigma^2}}$$

Taking the logarithms on both sides

$$ln(L(\mu, \sigma \mid x^{(1)}, x^{(2)}, \dots, x^{(n)})) \ = ln \ (\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x^{(1)}-\mu)^2}{2\sigma^2}} * \cdots * \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x^{(n)}-\mu)^2}{2\sigma^2}})$$

… apply log transormations…

$$= \frac{1}{2}\ln(2\pi) - \ln(\sigma) - \frac{\left(x^{(1)}-\mu\right)^2}{2\sigma^2} - \cdots - \frac{1}{2}\ln(2\pi) - \ln(\sigma) - \frac{\left(x^{(1)}-\mu\right)^2}{2\sigma^2}$$

$$\boxed{= \frac{n}{2}\ln(2\pi) - \text{n}\ln(\sigma) - \frac{\left(x^{(1)}-\mu\right)^2}{2\sigma^2} - \cdots - \frac{\left(x^{(1)}-\mu\right)^2}{2\sigma^2}}$$

Likelihood function of the Normal Distribution: Example (cont.)

$$L(\mu, \sigma \mid x^{(1)}, x^{(2)}, \dots, x^{(n)}) = \frac{n}{2}\ln(2\pi) - \mathrm{n}\ln(\sigma) - \frac{\left(x^{(1)}-\mu\right)^2}{2\sigma^2} - \cdots - \frac{\left(x^{(1)}-\mu\right)^2}{2\sigma^2}$$

Partial derivatives:

$$\frac{\partial \mathrm{L}(\mu,\sigma \mid x^{(1)},x^{(2)},\dots,x^{(n)})}{\partial \mu} = 0 - 0 + \frac{\left(x^{(1)}-\mu\right)}{\sigma^2} + \dots + \frac{\left(x^{(n)}-\mu\right)}{\sigma^2}$$

$$= \frac{1}{\sigma^2}[(x^{(1)} + \dots + x^{(n)}) - n\,\mu]$$

$$\frac{\partial \mathrm{L}(\mu,\sigma \mid x^{(1)},x^{(2)},\dots,x^{(n)})}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}[\left(x^{(1)} - \mu\right)^2 + \cdots + \left(x^{(n)} - \mu\right)^2]$$

Likelihood function of the Normal Distribution: Example (cont.)

$$\frac{\partial L(\mu,\sigma \mid x^{(1)},x^{(2)},\ldots,x^{(n)})}{\partial \mu} = \frac{1}{\sigma^2}[(x^{(1)}+\ldots+x^{(n)}) - n\,\mu] = 0$$

Multiply both sides by σ^2

$$\mu = \frac{(x^{(1)}+\ldots+x^{(n)})}{n}$$

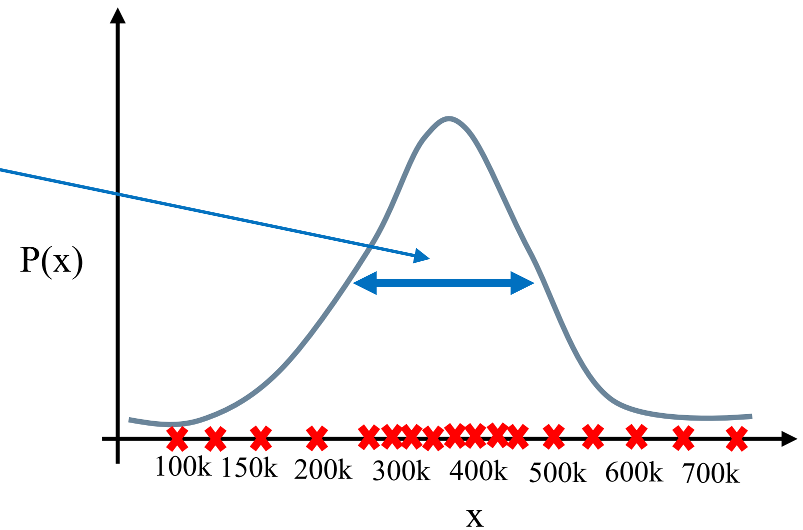
$$\frac{\partial L(\mu,\sigma \mid x^{(1)},x^{(2)},\ldots,x^{(n)})}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\left[\left(x^{(1)}-\mu\right)^2 + \cdots + \left(x^{(n)}-\mu\right)^2\right] = 0$$

$$\sigma = \sqrt{\frac{(x^{(1)}-\mu)^2 + \cdots + (x^{(n)}-\mu)^2}{n}}$$

Likelihood function of the Normal Distribution: Example (cont.)

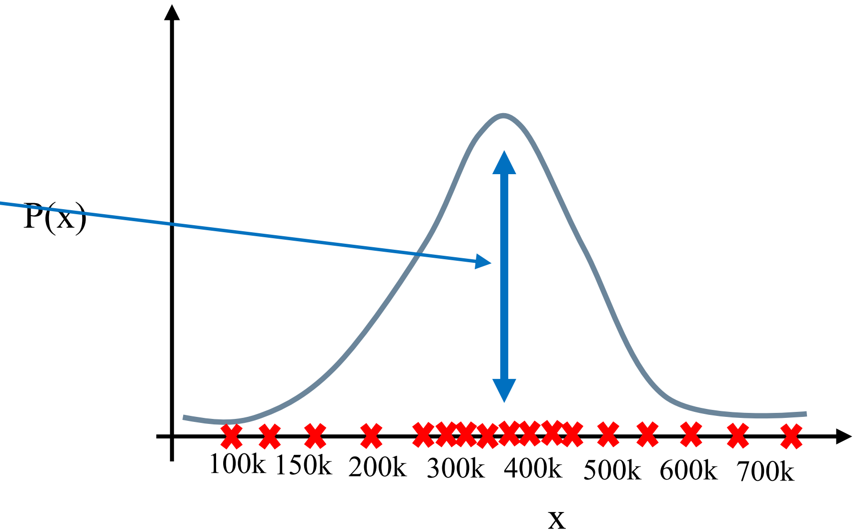
$$\mu = \frac{(x^{(1)} + \ldots + x^{(n)})}{n}$$

The mean of the data is the maximum likelihood estimate for where the center of the normal distribution



$$\sigma = \sqrt{\frac{(x^{(1)} - \mu)^2 + \cdots + (x^{(n)} - \mu)^2}{n}}$$

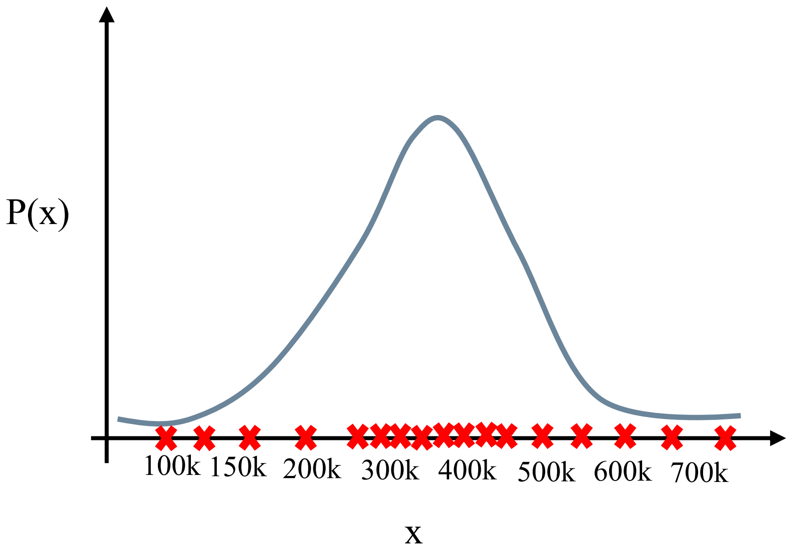
The standard deviation of the data is the maximum likelihood estimate how wide the normal distribution shoul be



Likelihood function of the Normal Distribution: Example (cont.)

$$\mu = \frac{(x^{(1)} + \ldots + x^{(n)})}{n}$$

$$\sigma = \sqrt{\frac{(x^{(1)} - \mu)^2 + \cdots + (x^{(n)} - \mu)^2}{n}}$$



These solutions may be obvious, but from maximum likelihood estimation, we prove that our intuition are correct.

- Thank you!