ECS171: Machine Learning

L10: Probabilistic methods

Instructor: Prof. Maike Sonnewald
TAs: Pu Sun & Devashree Kataria

MOCO Amsterdam garden

Midterm on tuesday.

How similar is to the practice?

Topic similarity
- Lectures 1-9 fair game

Recommended study plan
- Do practice exam
- Patch any holes exposed by practice exam
- Go through lectures 1-9 & write similar style questions
- Answer your self-written study guide

Similarity in the style of questioning
- ~ dozen short answer - MC, TF, write a couple sentences, do a short calculat
- ~ 3-4 calculation problems -

Activity 2 due Monday

Intended Learning Outcomes

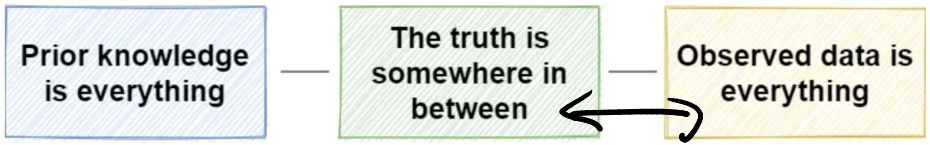
- Know and apply concepts of Bayesian probability
- Appreciate the connection between generative and discriminative classification
- For a simple regression case, appreciate how we fit a normal distribution using the likelihood
    - Insert numbers to compute the different model candidates

Bayesian Probability

- Bayes Theorem is a method to determine conditional probabilities
  - The probability of one event occurring given that another event has already occurred.
- Because a conditional probability includes additional conditions i.e. more data – it can contribute to more accurate results.

$P(E) = \frac{\#\,E}{\#\,\text{chances}}$ Frequentist

Points of View



$P(E)$: degree of belief ← "credence"

Thomas Bayes (1701-1761, UK)

Bayes was a minister

He never published his theorem

Bayes theorem

If our data (input vector), and our outcome (e.g. labels) is A

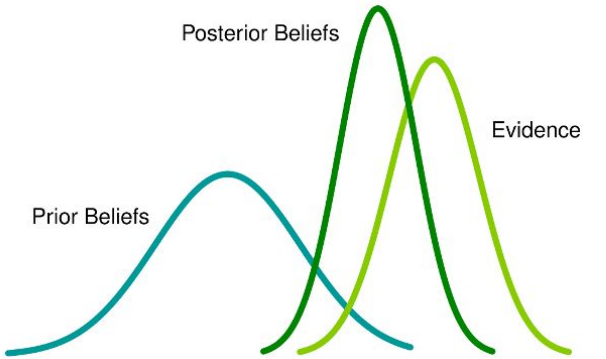
Conditional or posterior Probability: P(θ |data)
Probability of one (or more) event given the occurrence of another event, e.g. P(θ given data) or P(θ | data).

Likelihood: P(data|θ)

Marginal Probability or Prior: P(θ)
The probability of an event irrespective of the outcomes of other random variables

'Evidence': P(data)
What actually occurred

Joint Probability: P(θ, data) also written P(θ and data)
Probability of two (or more) simultaneous events



$$p(\theta \mid \text{data}) = \frac{p(\text{data} \mid \theta) \cdot p(\theta)}{p(\text{data})}$$

Posterior · Likelihood · Prior · Normalization

$$p(A \mid B) = \frac{p(B \mid A) \cdot p(A)}{p(B)}$$

Midterm Review

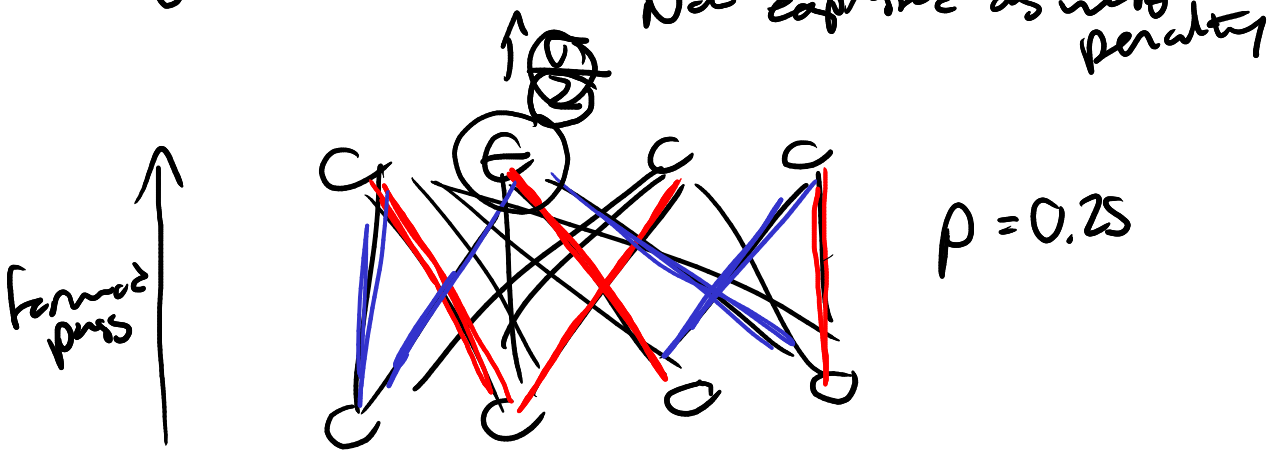
Q2

Extent of derivations

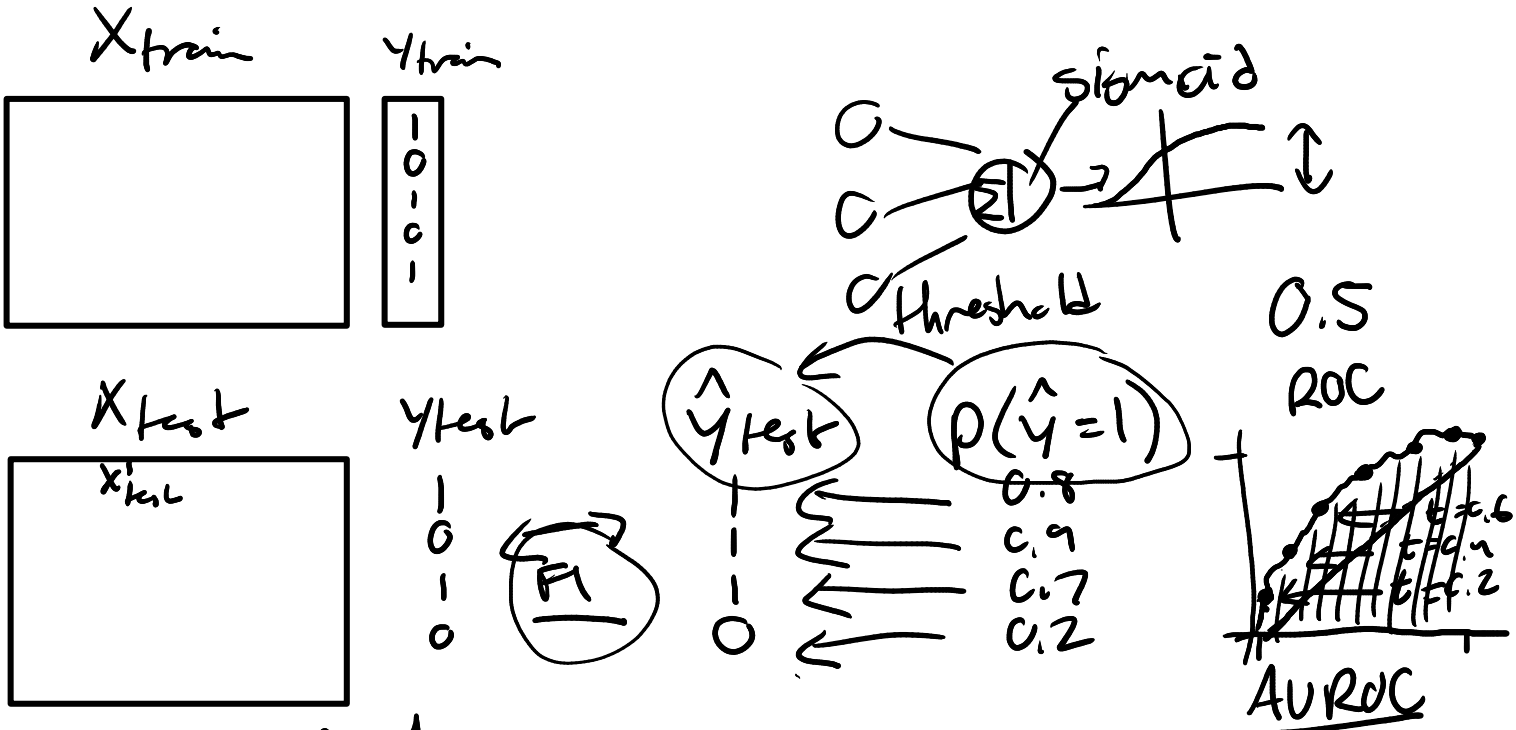
Dropout (not on real exam)

- Regularizer (Implicit)

Not expressed as weight penalty



Classification threshold



$$\hat{y}_{test} = f(x'_{test}) = \text{threshold}(\text{logit}(x'_{test}))$$

(neural network)

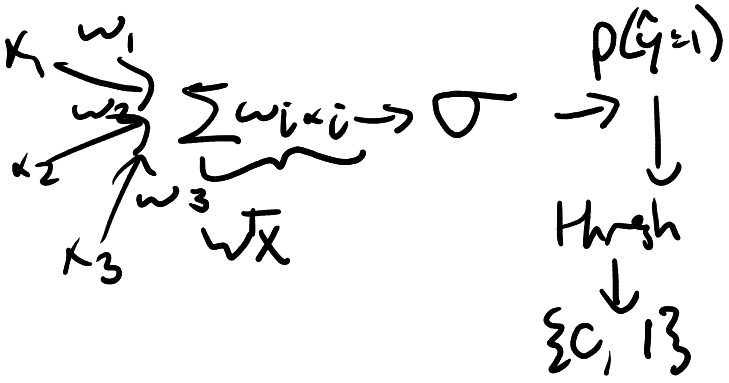
GS Practice Exam

Hypothesis function

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

Logistic regression

$\hat{y}$



prediction

$$\hat{y} = \sigma(w^T x)$$

hypothesis function

error

$$\varepsilon \sim N(0, \sigma^2)$$

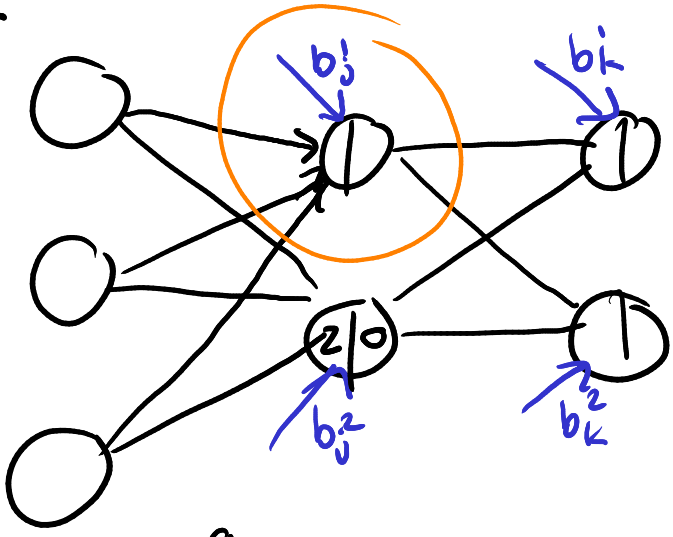
$$y = \sigma(w^T x) + \varepsilon$$

Q2

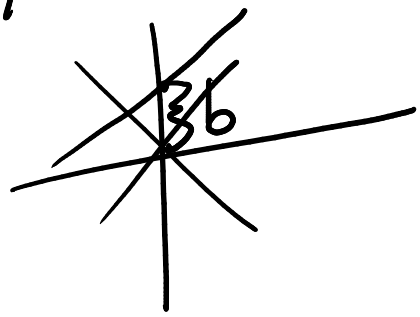
x_1 x_2 y $\hat{y}$

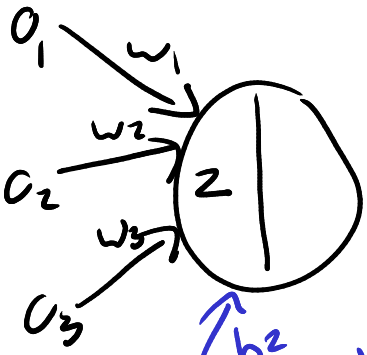
Which variance?
variance in $\hat{y}$

Q4



$$y = w^T x + b$$





$$z = \sum_i^3 w_i o_i + b_j^2$$

$$= \sum_{i=0}^3 w_i o_i$$

$b_j^2 = w_0$

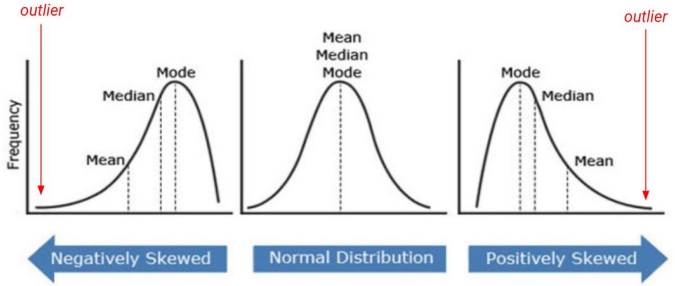
$o_0 = 1$

Recap: Statistics to Describe the Distribution of Data

- The *Mean* is the average of the data
- The *Median* is the value in the middle of the dataset (50% of the values smaller/larger or equal)
- The *Standard Deviation* (σ, square root of *variance*) measures the dispersion of data relative to its mean
- *Skewness* is a measure of the symmetry
- *Kurtosis* is the shape (tall, flat etc)
- **Note:** If a normal distribution is our 'model', we only need the mean and standard deviation to describe it

Moment number	Name	Measure of	Formula
1	Mean	Central tendency	$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$
2	Variance (Volatility)	Dispersion	$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N}$
3	Skewness	Symmetry (Positive or Negative)	$Skew = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{(X_i - \bar{X})}{\sigma} \right]^3$
4	Kurtosis	Shape (Tall or flat)	$Kurt = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{(X_i - \bar{X})}{\sigma} \right]^4$

Where X is a random variable having N observations (i = 1,2,…,N).

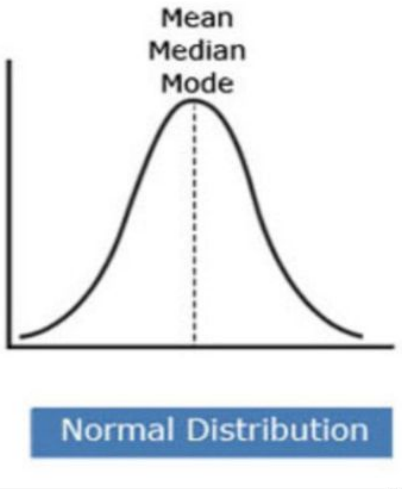


In negatively skewed data, the **mean is a smaller number** than the median and mode.	In a normal distribution, the **mean is the same number** as the median and mode.	In positively skewed data, the **mean is a larger number** than the median and mode.

The equation describing a normal distribution

- x = value of the variable or data being examined and f(x) the probability function
- μ = the mean
- σ = the standard deviation

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



Note: If a normal distribution is our ‘model’ we only need the mean and standard deviation to describe it!

We can treat classification in ‘probabilistic’ terms

(features)

$\mathbf{x}$

e.g:
- height
- weight
- color

(class label)

y

elefant

bear

Discriminative Classifiers: Direct mappings

- Learn mappings directly from the space of inputs X to class labels {0,1,2,...,K}
- For example:
  - Linear regression (as a classifier)
  - Neural Networks

Note ‘Discriminative’ defined as: making distinctions with accuracy

(features)

$\mathbf{x}$

e.g:
- height
- weight
- color

(class label)

y

elefant

bear

Discriminative Classifiers: Learn p(y|**x**) directly

- For example:
  - Logistic regression

(features)

$\mathbf{x}$

e.g:
- height
- weight
- color

(class label)

$p(y|\mathbf{x})$

elefant

bear

Generative Classifiers

- Build a model of how data for a class ‘looks like’
- Generative classifiers try to mode p(**x**|y)
- Classification via Bayes rule
  - Called Bayes Classifiers

(prob. of features given label)

$p(\mathbf{x}|y)$

e.g:
- height
- weight
- color

(class label)

y

elefant

bear

Approaches to classification: Generative vs Discriminative

- Discriminative classifiers estimate parameters of decision boundary/class separator directly from labeled examples
  - Learn p(y |**x**) directly (logistic regression models)
  - Learn mappings from inputs to classes (least-squares, neural nets)
- Generative approach: model the distribution of inputs characteristic of the class (Bayes classifier)
  - Build a model of p(**x**|y)
  - Apply Bayes Rule

Bayes Classifier

- Aim to diagnose whether patient has diabetes:
  - Classify into one of two classes (yes C=1; no C=0)
  - Run a number of tests (d) on subjects, get **x** for each patient
- Compute class given a patient's result: $\mathbf{x} = [x_1, x_2, \ldots, x_d]^T$
  - Use Bayes Rule:

$$p(C|\mathbf{x}) = \frac{p(\mathbf{x}|C)p(C)}{p(\mathbf{x})}$$

  - Put differently:

$$\text{posterior} = \frac{\text{Class likelihood} \times \text{prior}}{\text{Evidence}}$$

- How can we compute p(x) for the two class case?

$$p(\mathbf{x}) = p(\mathbf{x}|C=0)p(C=0) + p(\mathbf{x}|C=1)p(C=1)$$

- To compute p(C|**x**) we still need: p(**x**|C) and p(C)

...**but let's do a *simple regression* analysis first for building intuition…**

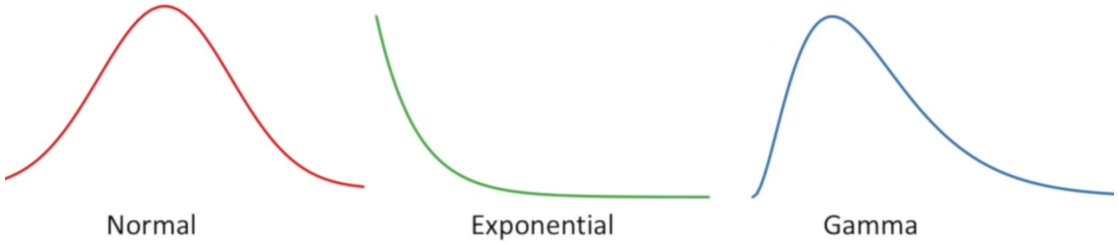
Recall L4: We can use the maximum likelihood estimation

- To determine the function that maximises the likelihood of having a good model use the method: Maximum likelihood estimation (MLE)
- First: We must choose a probability distribution we believe is a good fit
  - Normal distribution is a good place to start

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

Recall L4: We can use the maximum likelihood estimation

- To determine the function that maximises the likelihood of having a good model use the method: Maximum likelihood estimation (MLE)
- First: We must choose a probability distribution p(x|C) we believe is a good fit
  - Normal distribution is a good place to start



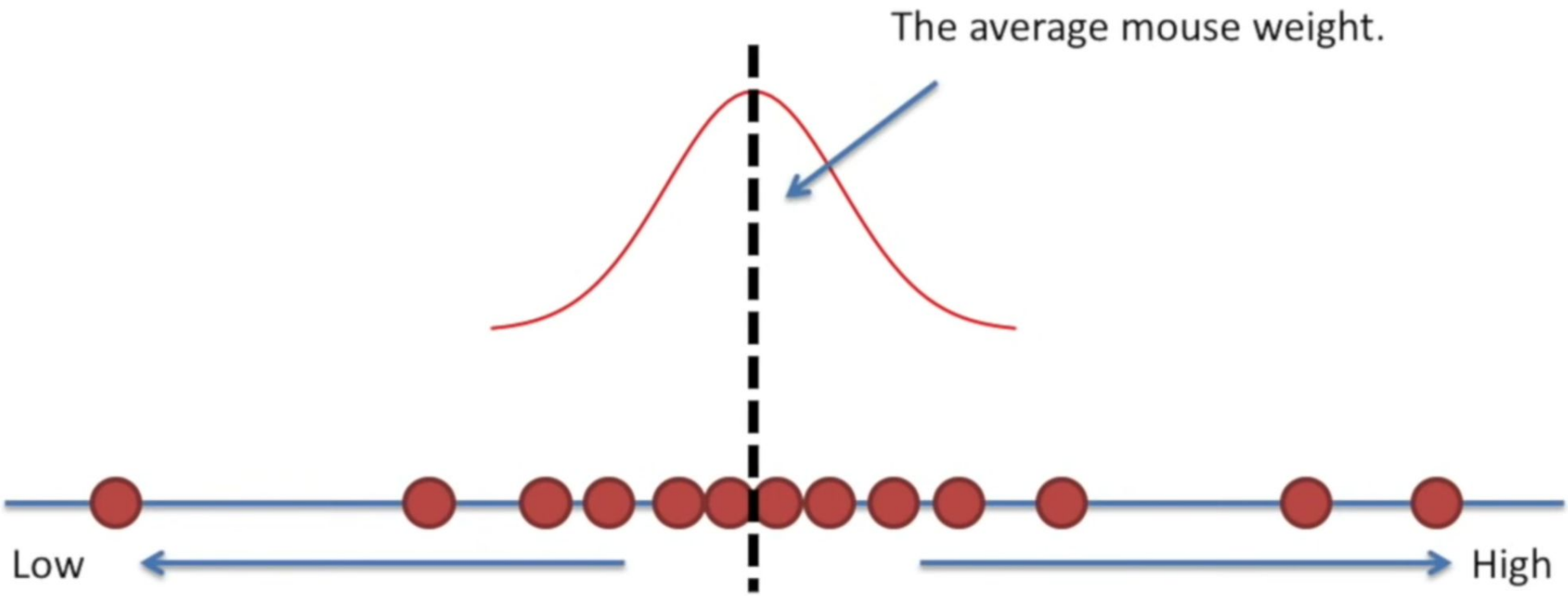
Maximum Likelihood Estimation

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

- Fitting a model to a series of mouse weights
- Let's try a normal distribution

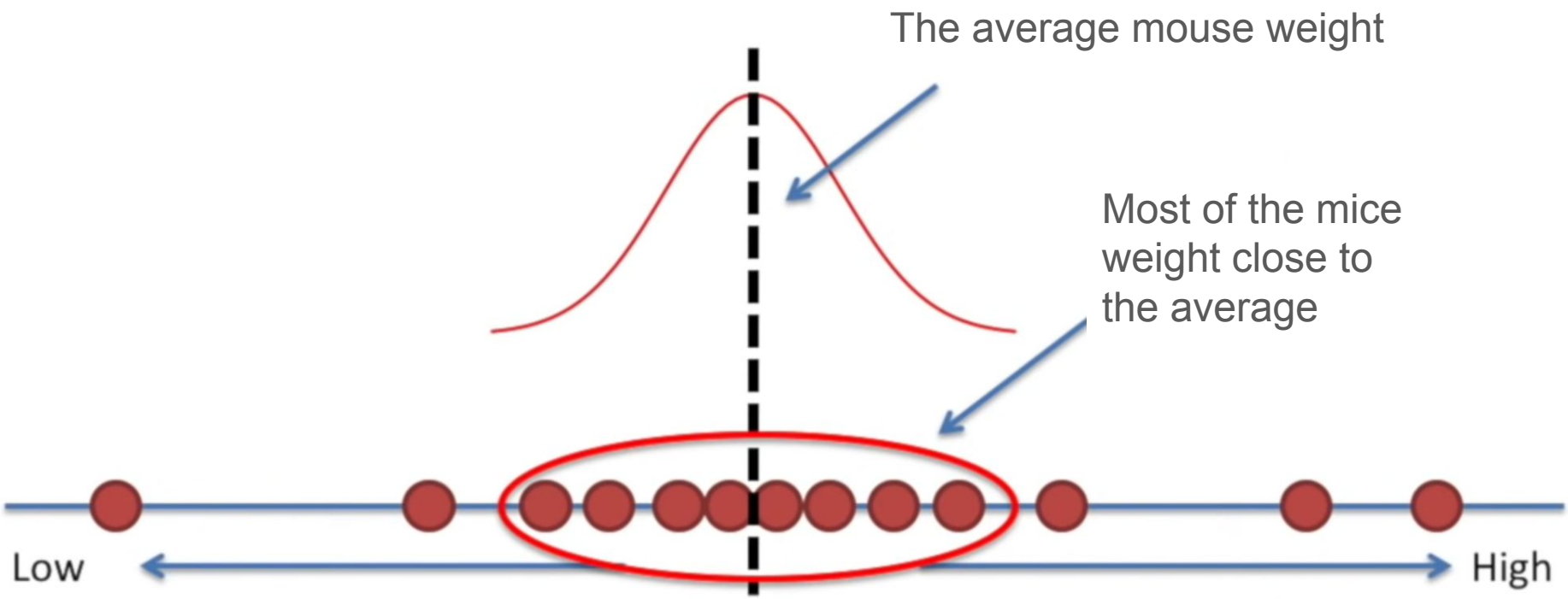
Maximum Likelihood Estimation

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



Maximum Likelihood Estimation

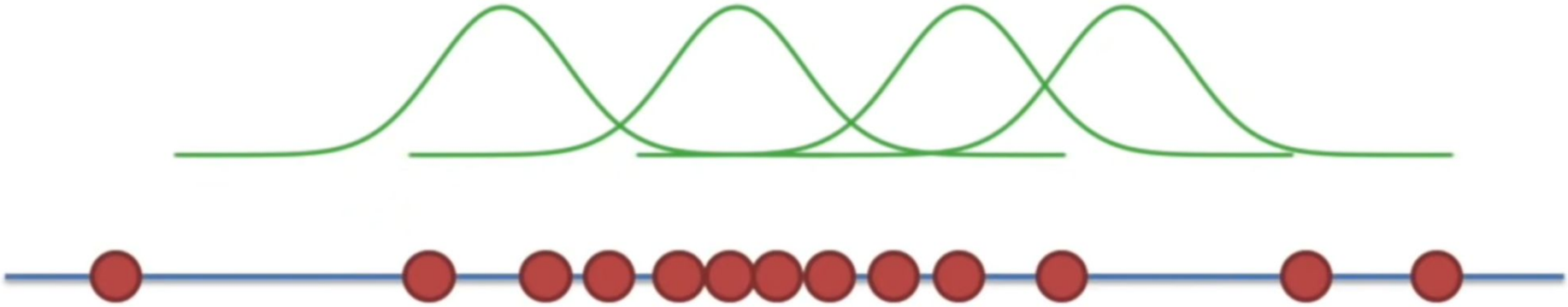
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Maximum Likelihood Estimation

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

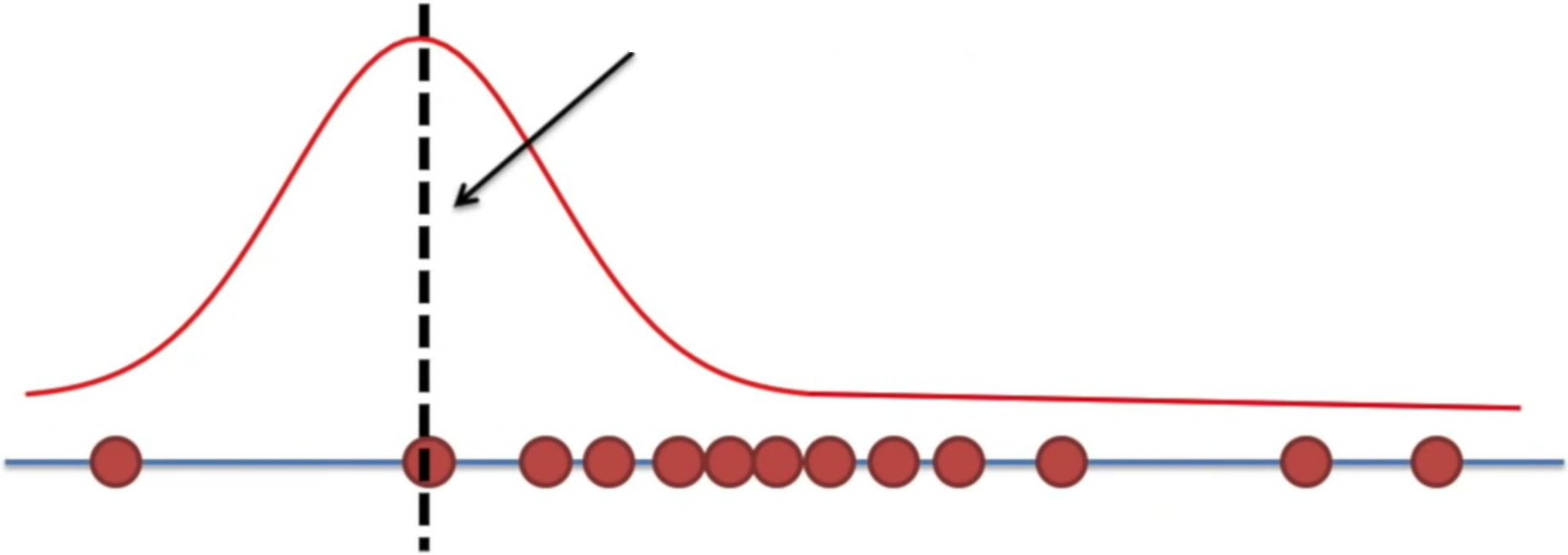
Once shape is determined, we must determine the optimal location that maximises the probability of being similar to the observations

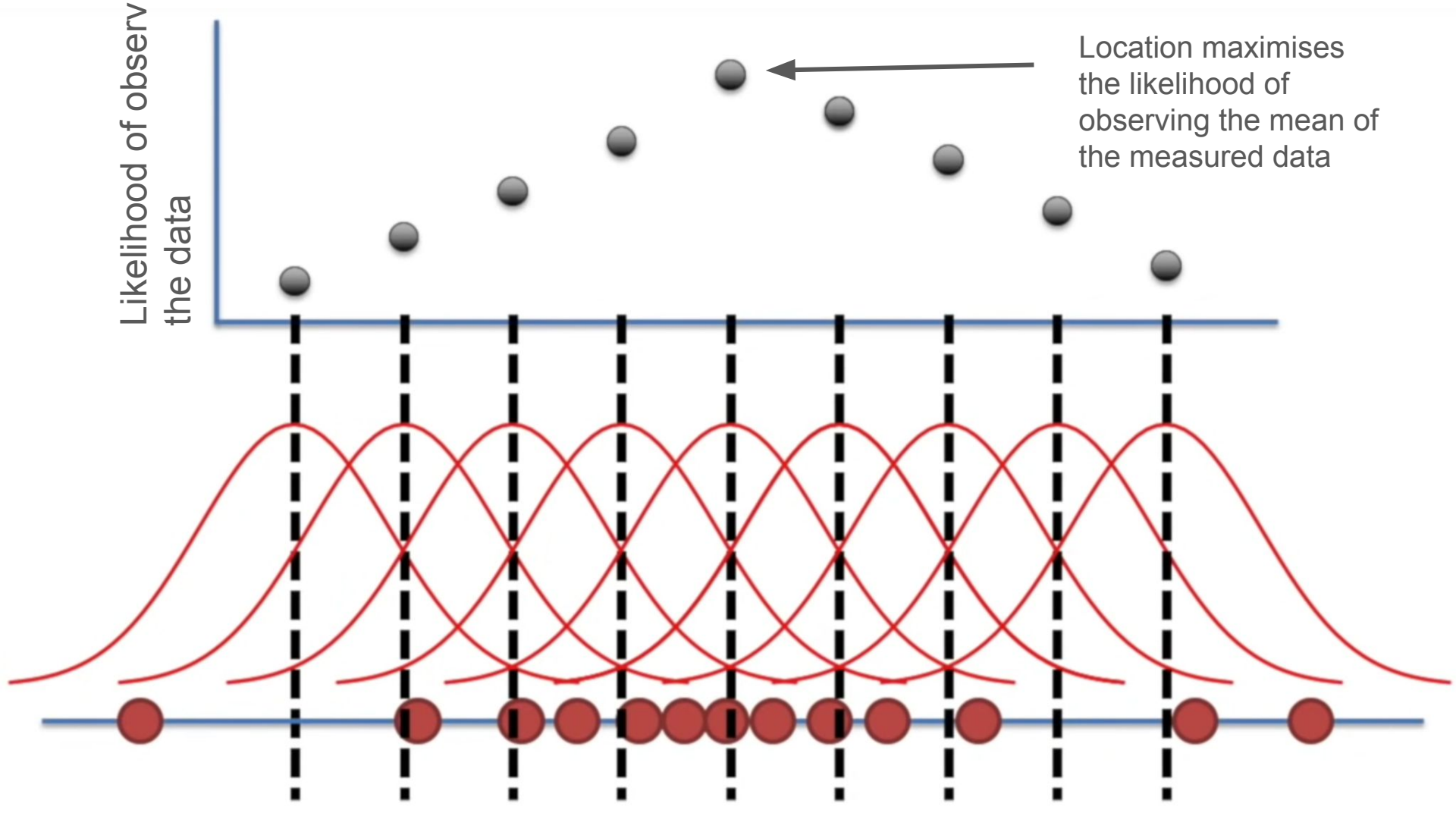


Maximum Likelihood Estimation

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

Most of the measured values should be near the average

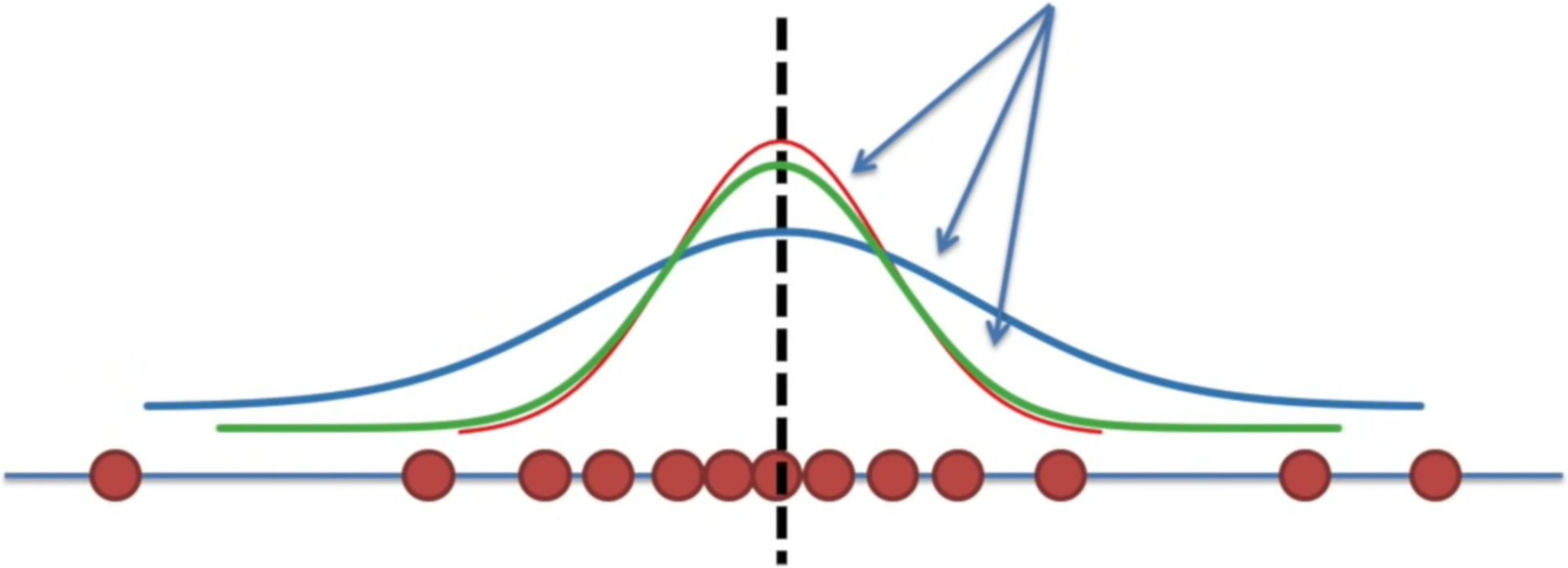


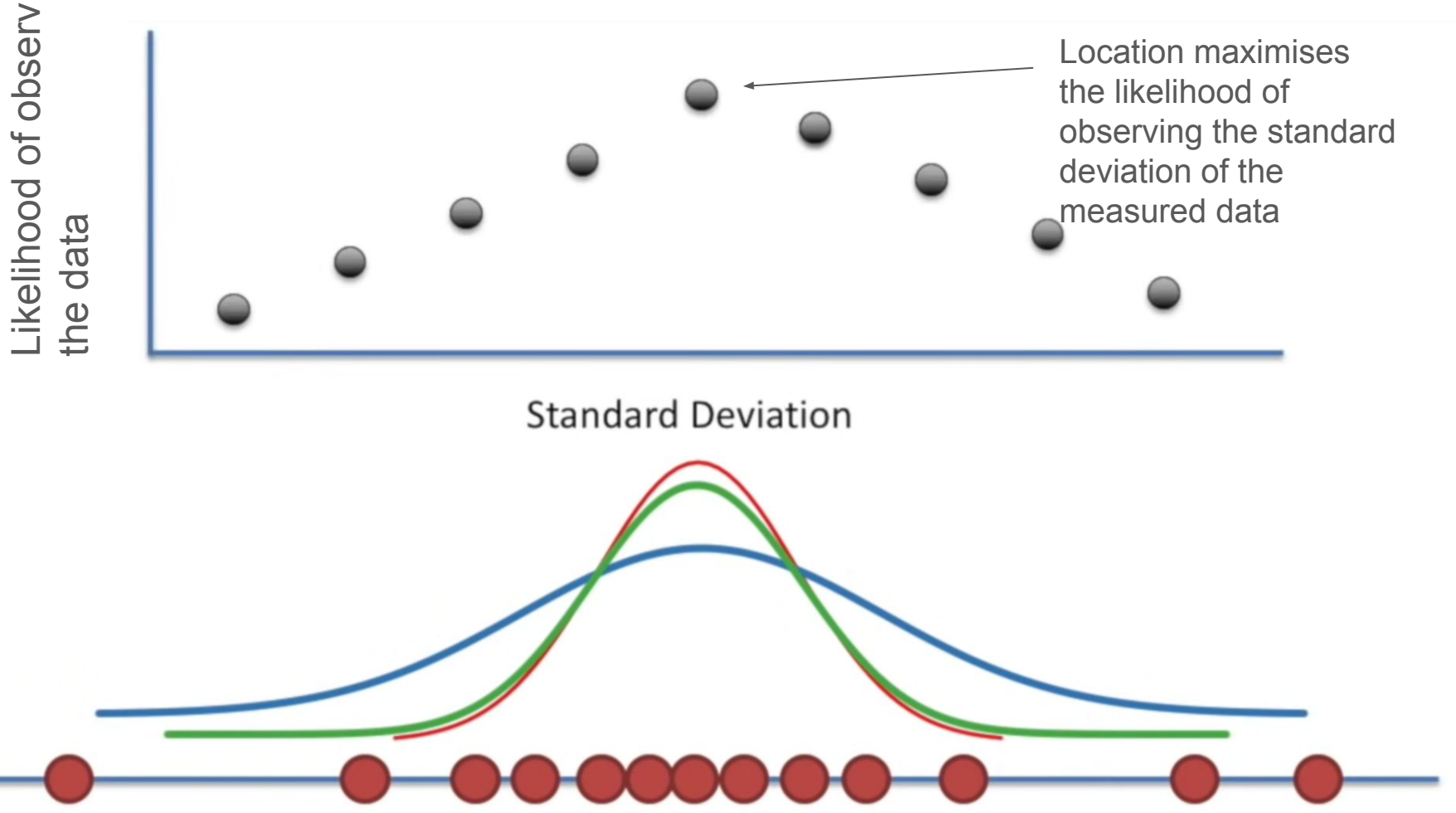

Likelihood of observ
the data
Location maximises
the likelihood of
observing the mean of
the measured data

Maximum Likelihood Estimation

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

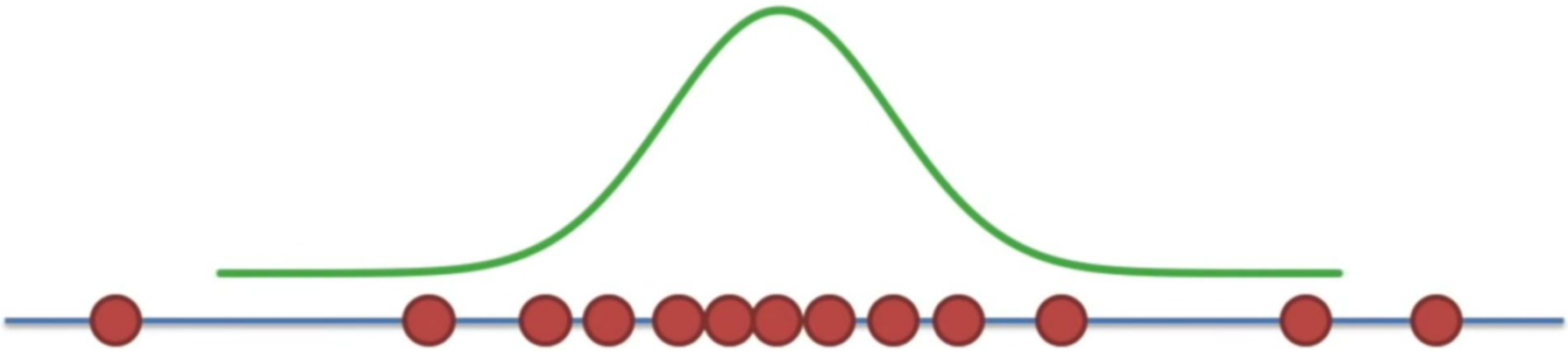
Now we need the ‘maximum likelihood’ of the standard deviation




Likelihood of observ
the data
Location maximises the likelihood of observing the standard deviation of the measured data
Standard Deviation

Maximum Likelihood Estimation

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



Maximum Likelihood Estimation

We now have the function: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

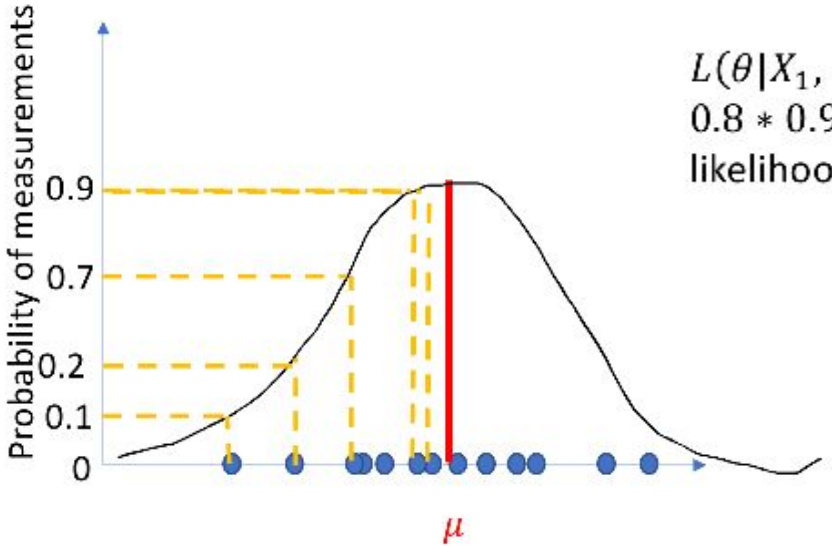
And we want to estimate what a given sample would result in varying the mean and standard deviation:

$$
\begin{aligned}
L(\theta|X_1, \ldots, X_n) &= L(\mu, \sigma|X_1, \ldots, X_n) \\
&= L(\mu, \sigma|X_i) \\
&= L(\mu, \sigma|X_1) \times \ldots \times L(\mu, \sigma|X_n) \\
&= \prod_{i=1}^{n} L(\mu, \sigma|X_i)
\end{aligned}
$$

Maximum Likelihood Estimation

Given samples (blue dots), some mean and standard deviation, insert the data

Determine the likelihood of the entire dataset



$L(\theta|X_1, \ldots, X_n) = L(\mu, \sigma|X_1, \ldots, X_n) = 0.1 * 0.2 * 0.7 * 0.7 * 0.8 * 0.9 * 0.9 * 0.9 * 0.9 * 0.9 * 0.9 * 0.3 * 0.1$= overall likelihood of the entire dataset.

Fitting a Normal Distribution fit

1. Start with a statistical model
   a. such as a normal distribution—the parameters would be μ and σ
2. Construct the likelihood function
3. Maximize the Likelihood function
4. Find the estimates

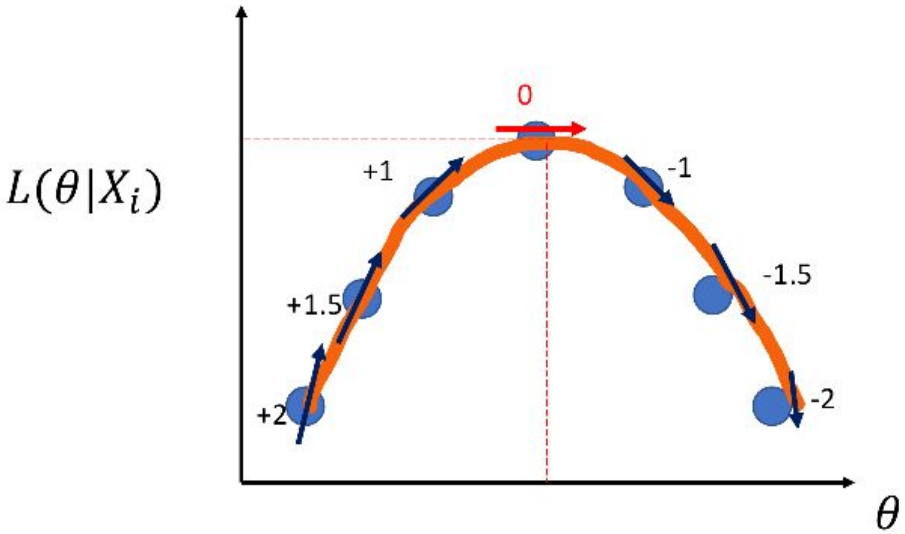
Recall: We are using a simple linear regression example (should be familiar)

Linear regression model parameters can be estimated using negative log likelihood function from MLE.

The negative log likelihood function can now be used to derive the least squares solution

Maximum Likelihood Estimation

Goal: Find the best fit where the likelihood of θ given the measurements is maximum.



Converting Linear Regression to a Probability Density Function (pdf)

Given a fixed, non-random sample X_1 , X_2,...,X_n find the best distribution that fits Y_1, Y_2,...,Y_n using a normal distribution, where a linear regression function is of the form:

$$y = f(x) + \epsilon$$

$$f(x) = \alpha + \beta x$$

$$\epsilon \sim N(0, \sigma^2)$$

For a fixed sample X_i the distribution of Y_i is equal to:

$$N(f(X_i), \sigma^2)$$

Recall, for our normal distribution we have have:

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

Converting Linear Regression to a Probability Density Function (pdf)

Recall, for our normal distribution we have have: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$$y_i = f(x_i) + \epsilon$$

$$f(x_i) = \alpha + \beta x_i$$

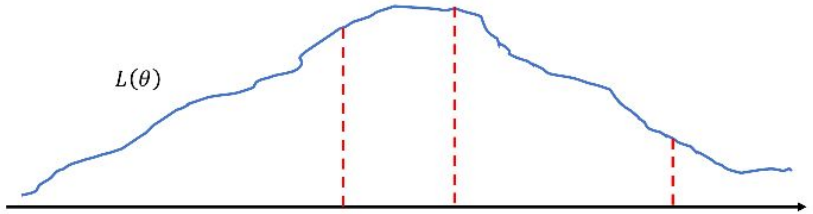
$$\epsilon_i = y_i - f(x_i) = y_i - \mathbf{w}x_i$$

Setting the mean to zero (μ = 0) we have:

$$f_{pdf(\epsilon_i)} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - f(x_i))^2}{2\sigma^2}}$$

Using the likelihood function

Here, θ is our unknown parameter of L(θ)



We now have the different X_1, X2,..,X_n arriving at:

$$L(\theta|X_1, \dots, X_n) = f(X_1, \dots, X_n|\ \theta)$$
$$= f(X_i|\ \theta)$$
$$= f(X_1|\ \theta) \dots f(X_n|\ \theta)$$
$$= \prod_{i=1}^{n} f(X_i|\ \theta)$$

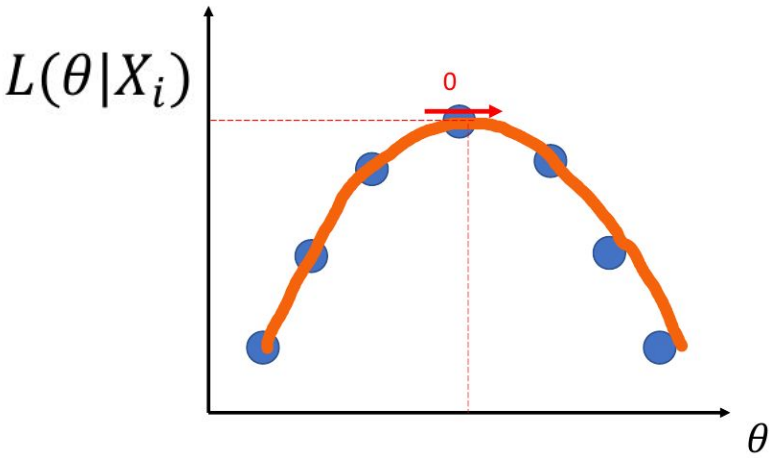
Which we now write as the Negative Log Likelihood:

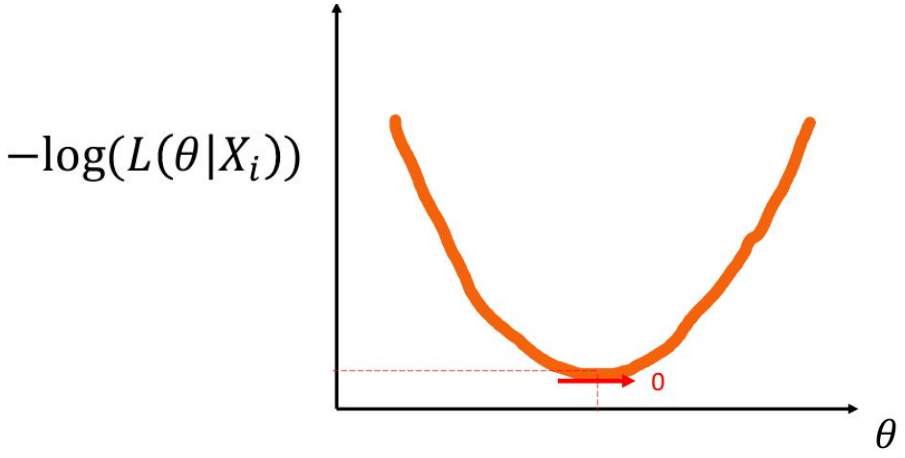
$$L(\theta|X_1, \dots, X_n) \cong \ - \ log\ \ L(\theta|X_1, \dots, X_n)$$

We can now look for the maximum/minimum

Using methods of maximization we can find the optimal value

Which with the minus sign becomes a maximisation for the Negative Log Likelihood





MLE continued

Recall, as the likelihood is usually a very small value, we take the negative log:

$$L(\theta|X_1, \dots, X_n) \cong \; - \; log \;\; L(\theta|X_1, \dots, X_n)$$

Arriving at the Negative Log Likelihood (NNL) we can carry along what we saw in

$$L(\theta) \triangleq \log L(\theta|X) = \sum_{i=1}^{N} \log L(\theta \,|\, X_i)$$

$$= \sum_{i=1}^{N} \log[\frac{1}{\sqrt{2\pi\sigma^2}} \times \;\; e^{-\frac{(Y_i - w^T X_i)^2}{2\sigma^2}}]$$

$$= -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(Y_i - w^T X_i)^2$$

MLE continued

Negative Log Likelihood (NLL)

Residual Sum of Squares (RSS)

$$NLL(\theta) \triangleq -\log L(\theta|X) = -\sum_{i=1}^{N} \log L(\theta|X_i)$$

$$RSS(w) \triangleq \sum_{i=1}^{N}(Y_i - w^T X_i)^2 = ||\epsilon||^2$$

sum of squared errors (SSE)

squared norm of residual errors.