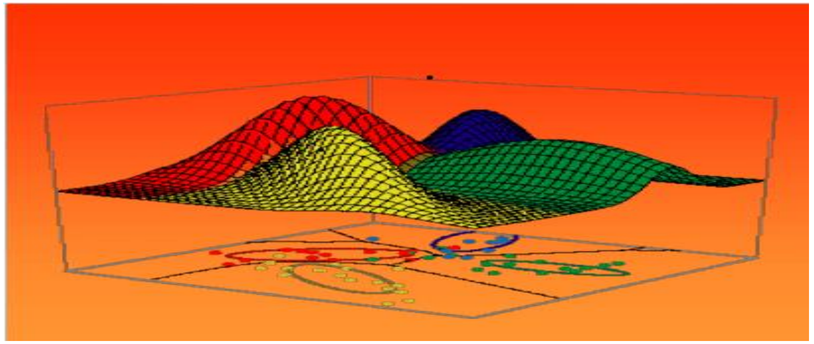


Pattern Recognition (CSE4213)

Faisal Muhammad Shah
Associate Professor,Dept Of CSE, AUST

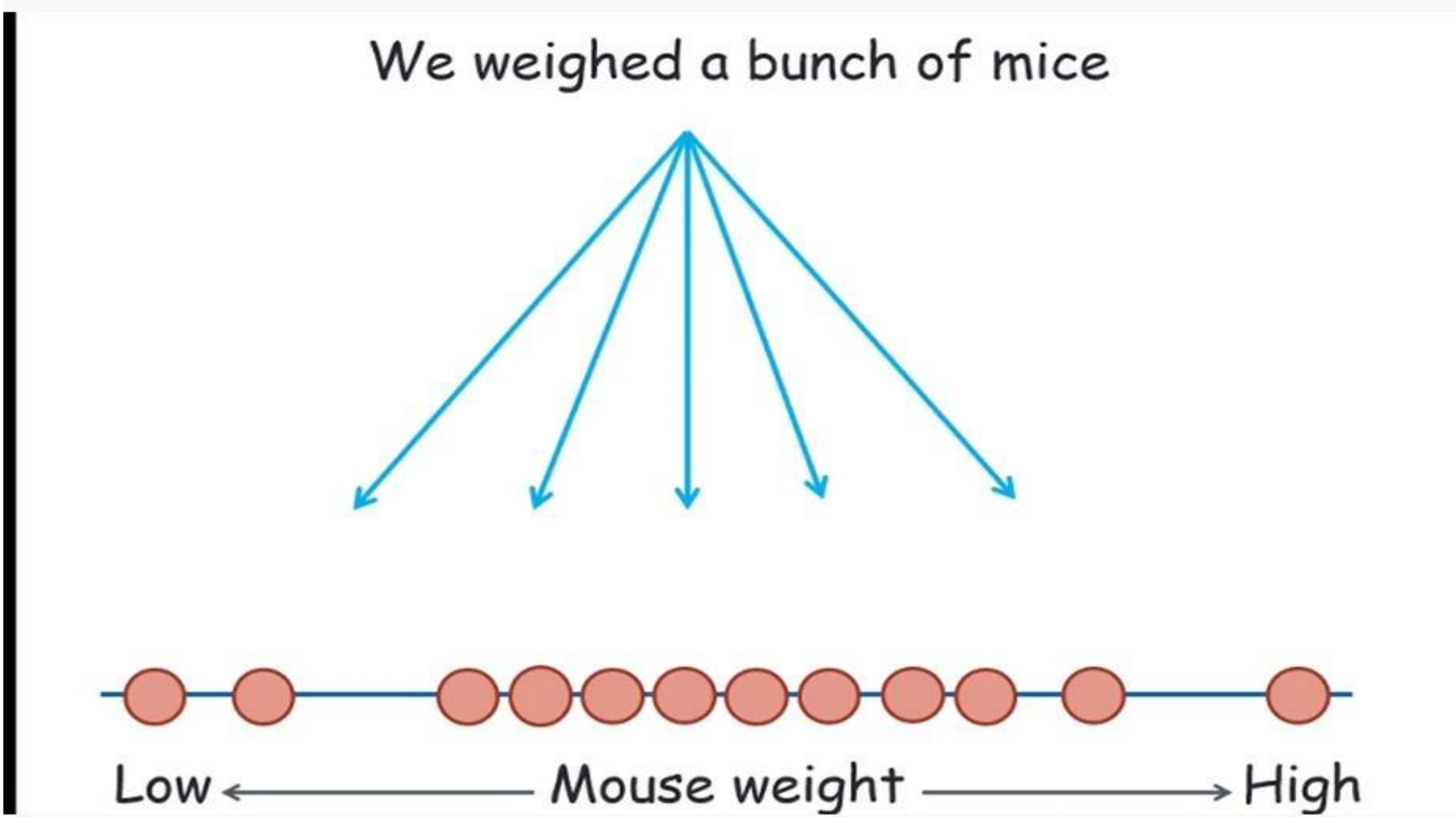
- **Parameter Estimation** - Chapter 3 (Duda et al.)
  - ***Maximum Likelihood Estimation***

Introduction

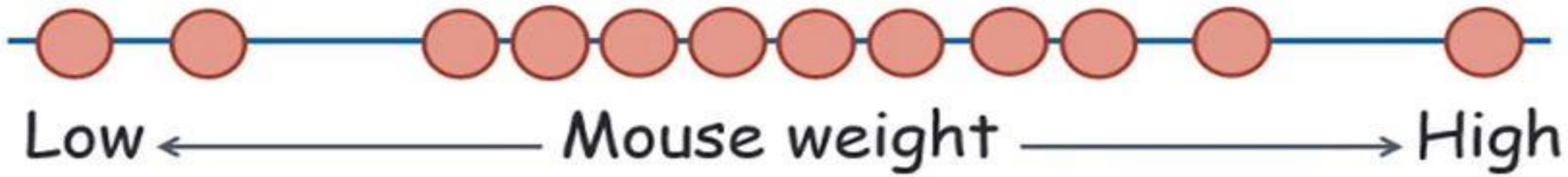
- In Chap 2, we saw how we could design an optimal classifier if we knew the **prior $P(\omega_i)$** and the likelihood/**class-conditional densities $p(x|\omega_i)$.**

- Unfortunately, in pattern recognition applications we rarely if ever have this kind of complete knowledge about the probabilistic structure of the problem.

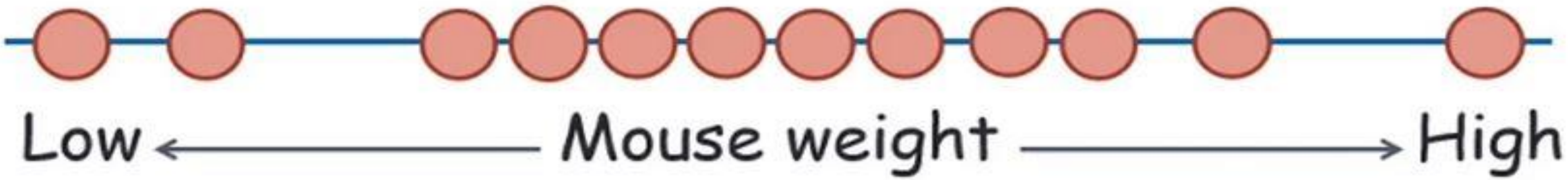
- The problem, then, is to find some way to use this information to design or data train the classifier.


We weighed a bunch of mice
Low
Mouse weight
High

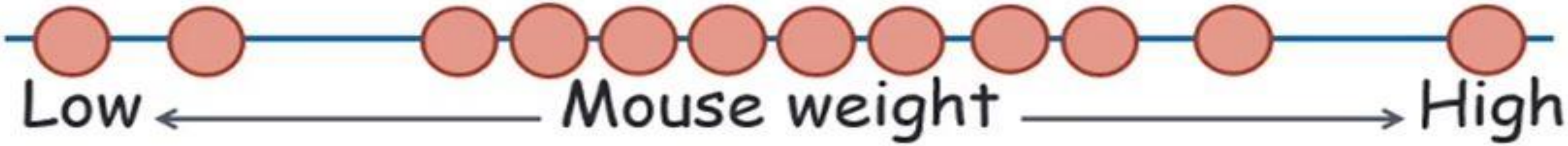
The goal of maximum likelihood is to find the optimal way to fit a distribution to the data



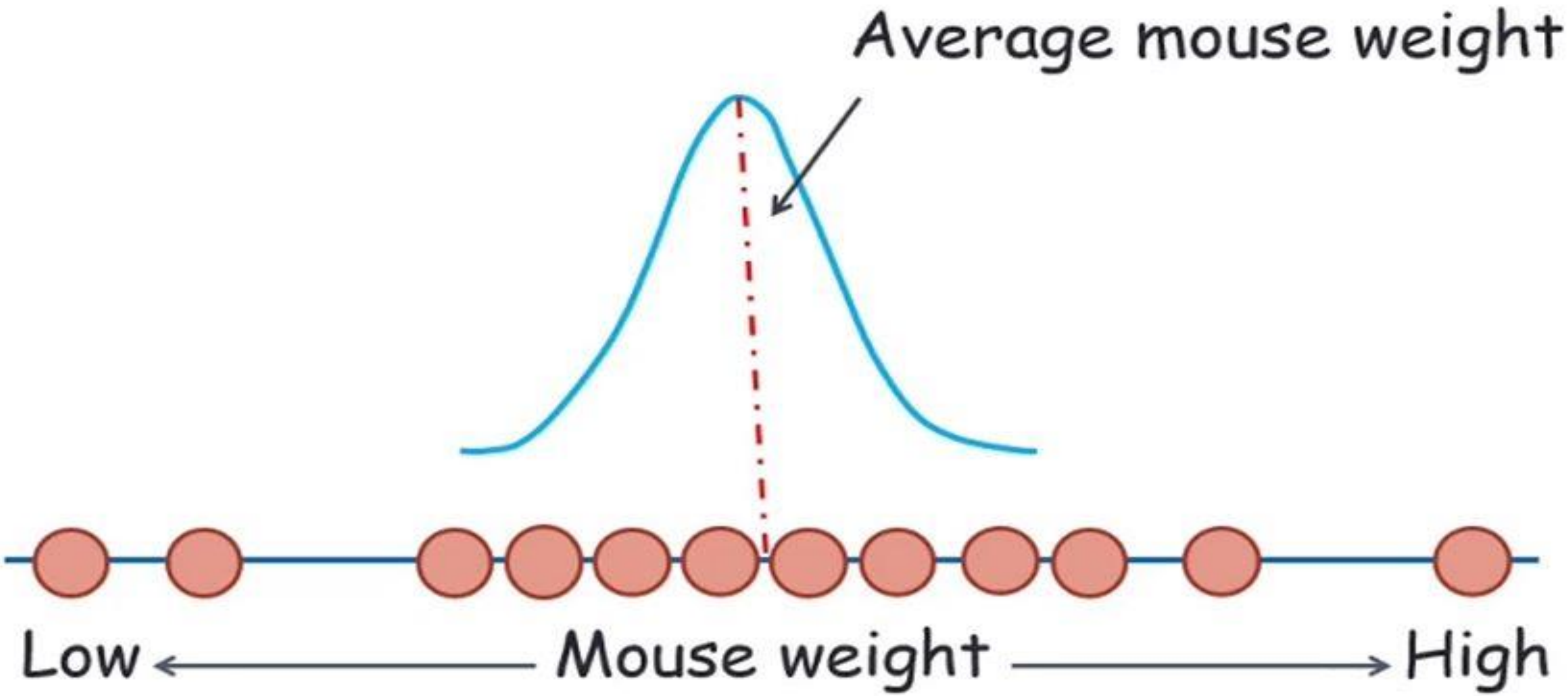
There are a lot of distributions for different types of data



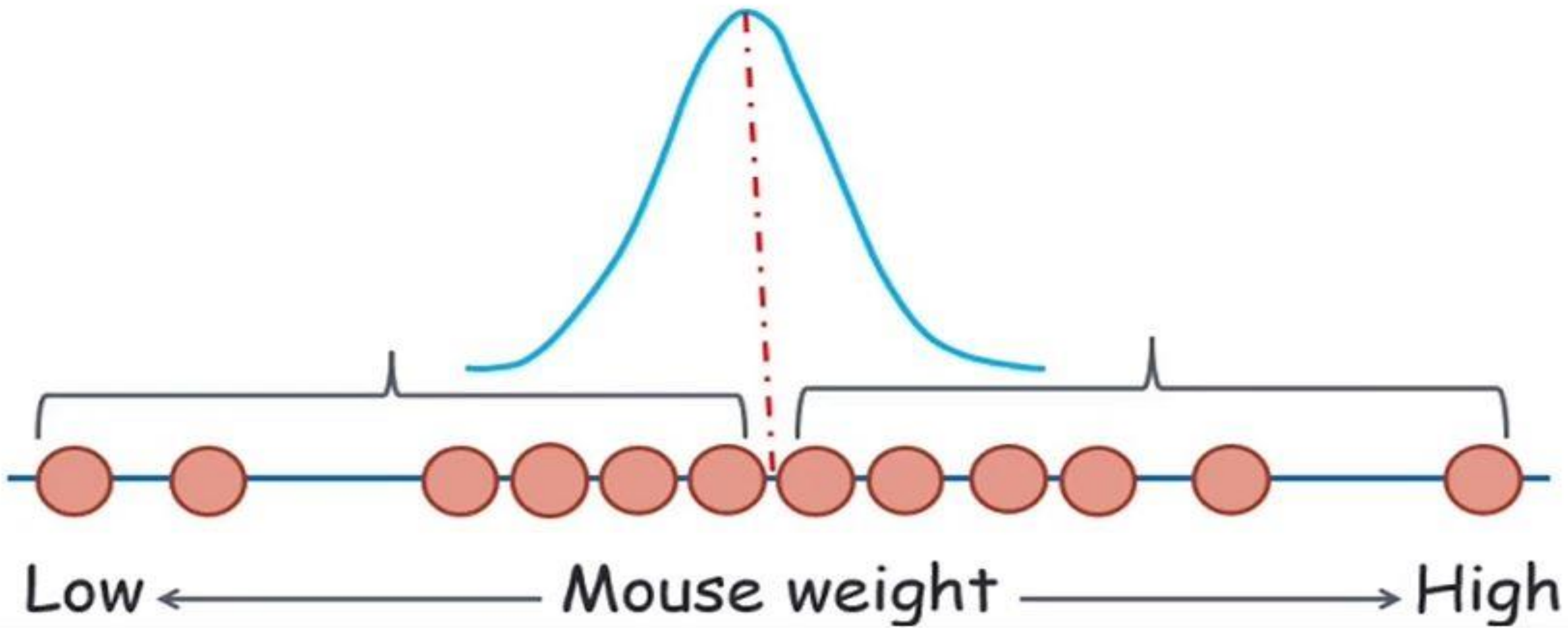
The reason we want to fit a distribution to our data is it can be easier to work with and it is also more general- it applies to every experiment of the same type

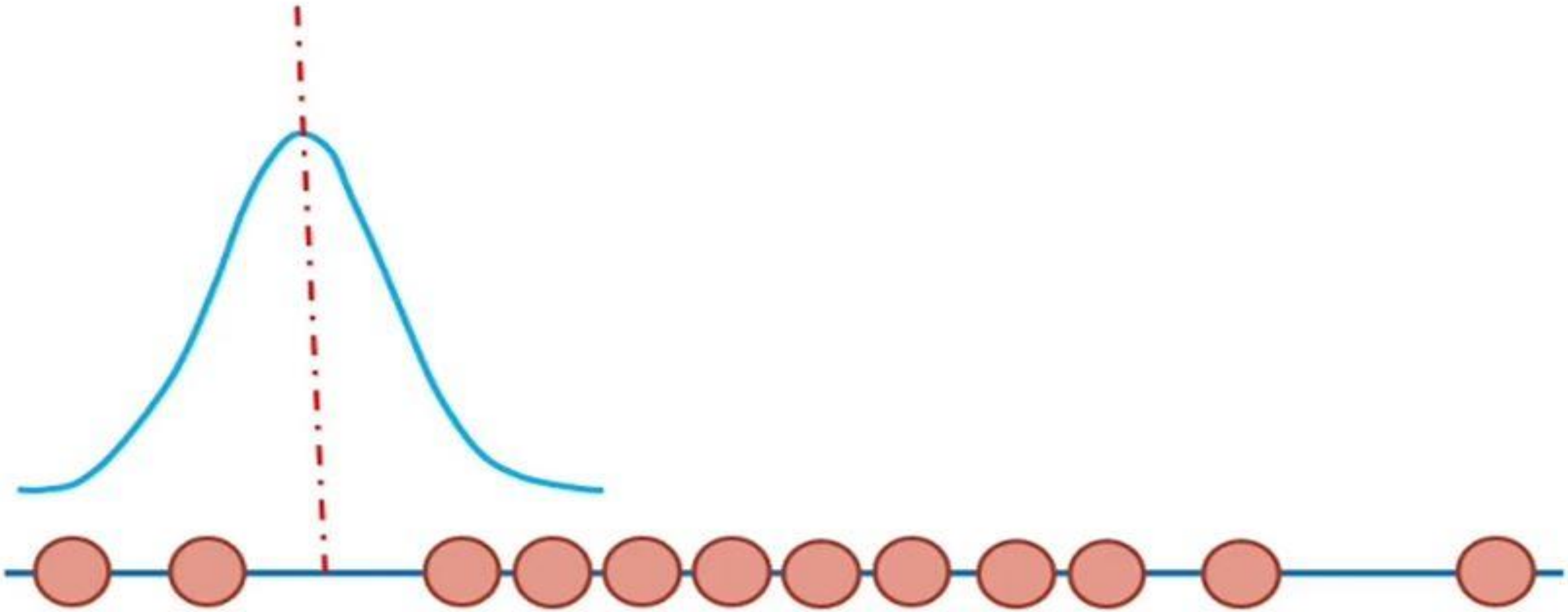


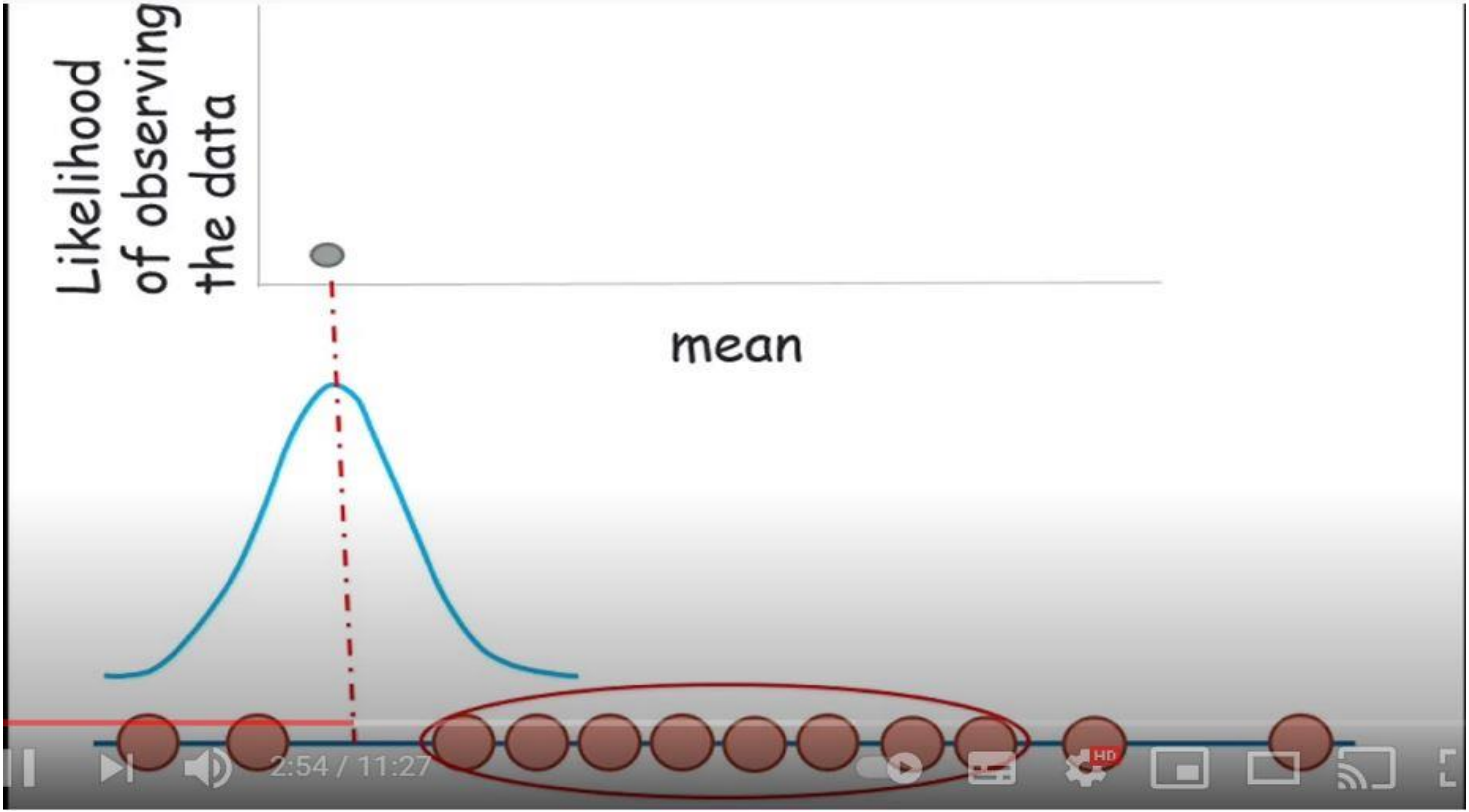
That means we expect most of the measurements (weights) are close to the mean.

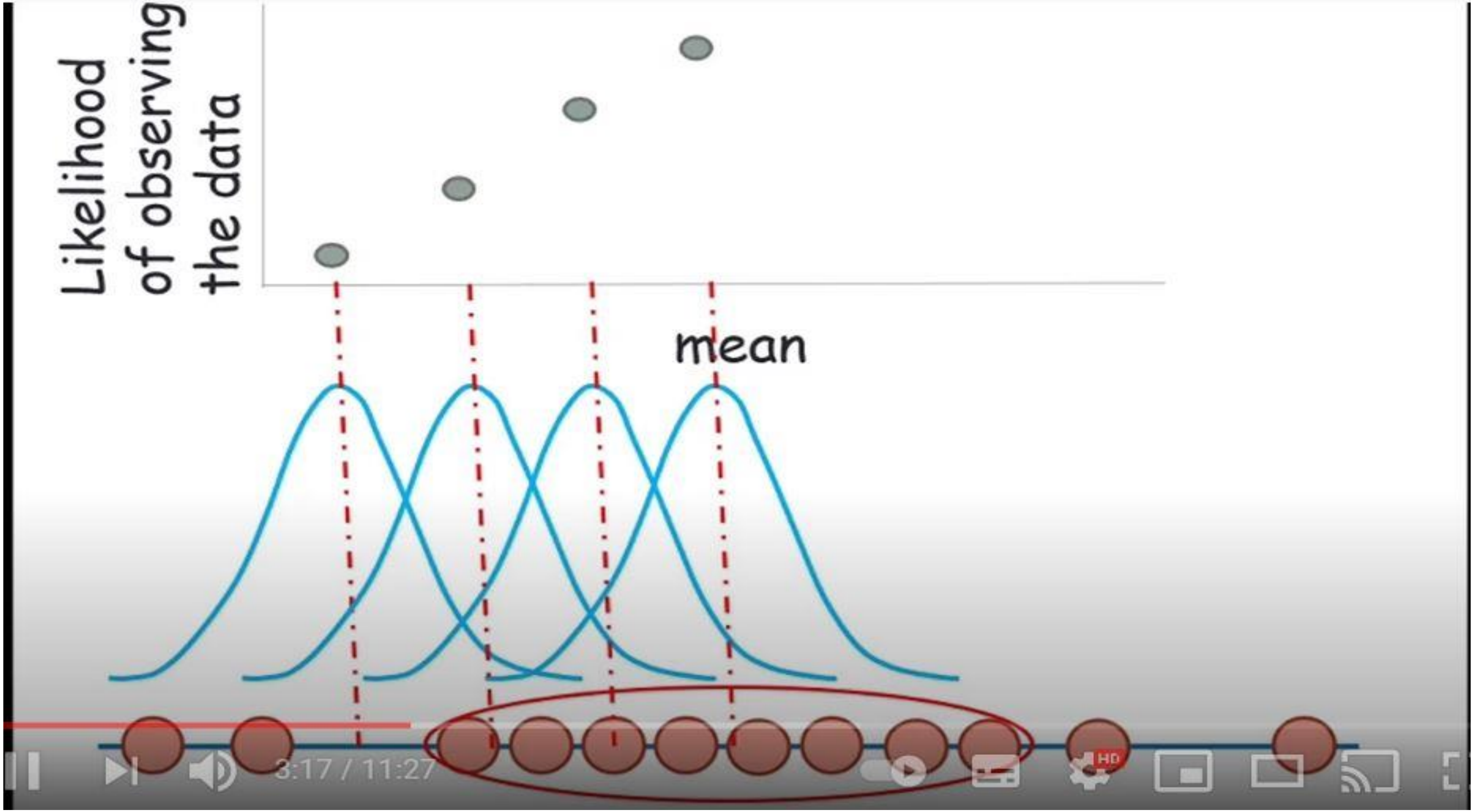


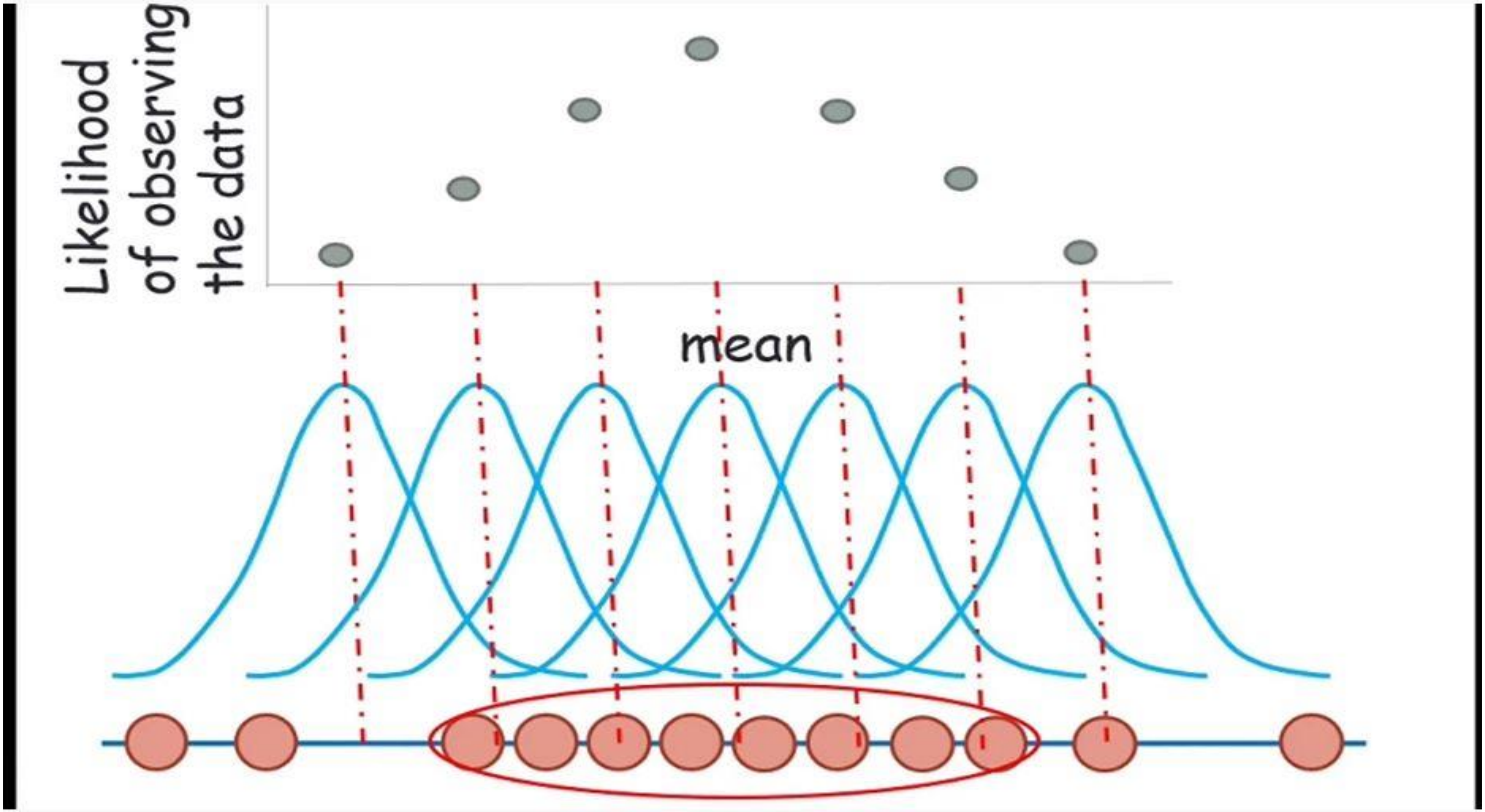
Relatively symmetrical around the mean.

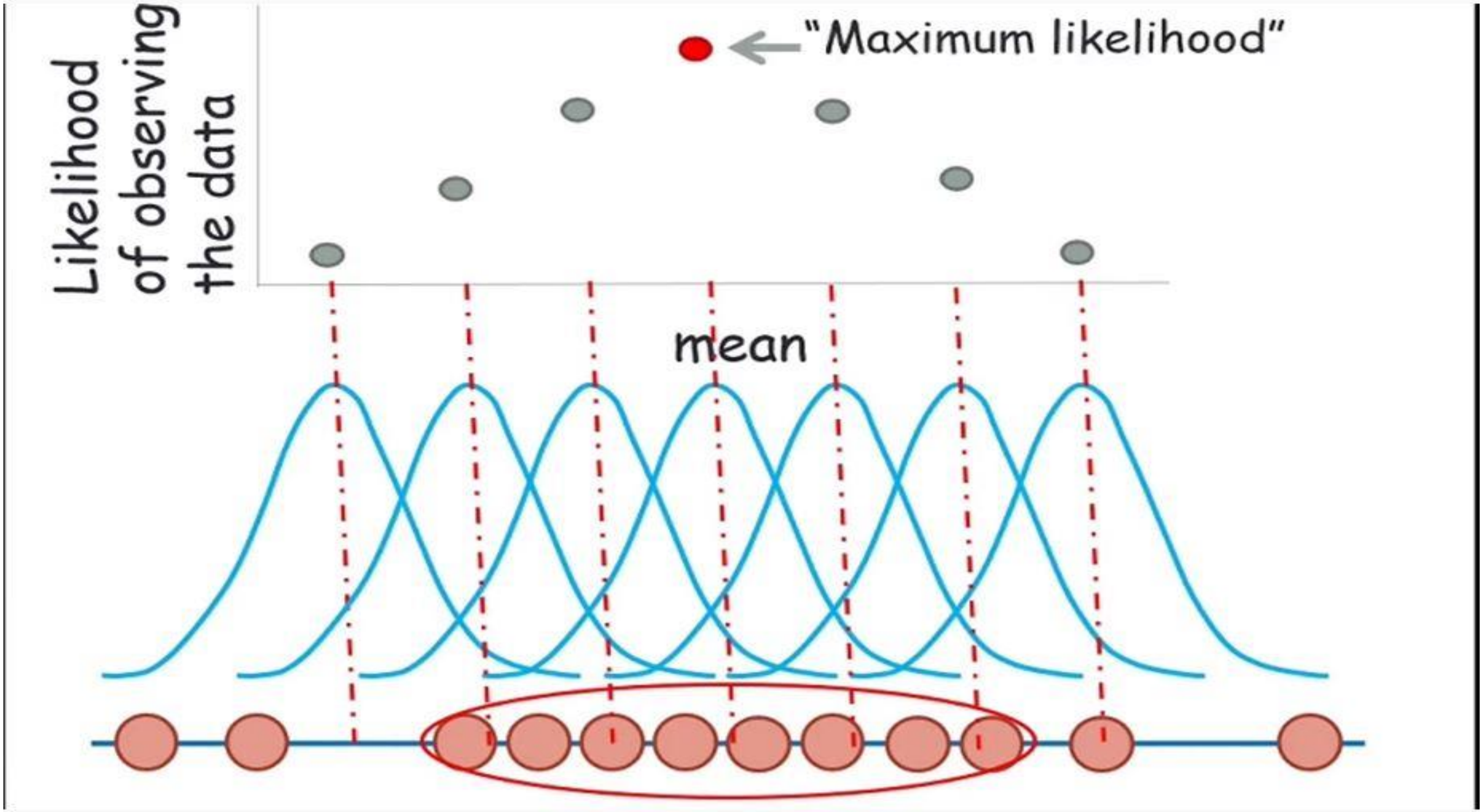





Likelihood of observing the data
mean
2:54 / 11:27
HD


Likelihood of observing the data
mean
3:17 / 11:27


Likelihood
of observing
the data
mean


Likelihood of observing the data
"Maximum likelihood"
mean

Introduction

- One approach to this problem is to use the samples to estimate the unknown probabilities and probability densities, and to use the resulting estimates as if they were the true values.

- In typical supervised pattern classification problems, the estimation of the prior probabilities presents no serious difficulties.

- However, estimation of the class-conditional densities is quite another matter.

- Serious problems arise when the dimensionality of the feature vector **x** is large.

Introduction

- The severity of these problems can be reduced significantly

  - If we know the number of parameters in advance and our general knowledge about the problem permits us to parameterize the conditional densities.

- Suppose, for example, that we can reasonably assume that $p(\mathbf{x}/\omega_i)$ is a normal density with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$, although we do not know the exact values of these quantities.

- This knowledge simplifies the problem from one of estimating an unknown *function* $p(\mathbf{x}/\omega_i)$ to one of estimating the *parameters* $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$

Introduction

- We shall consider two common and reasonable parameter estimation procedures,

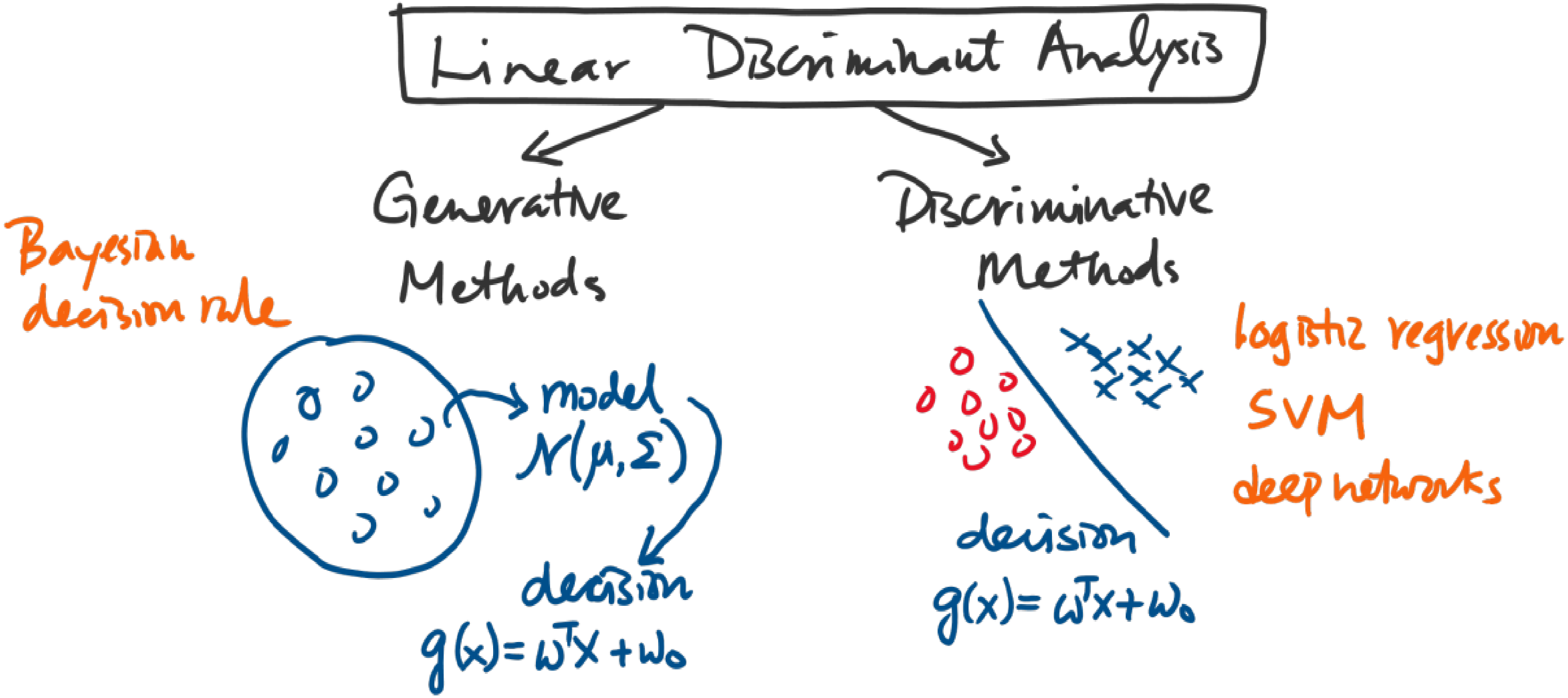
  ***-Maximum Likelihood Estimation*(MLE).**

  ***-Bayesian* Estimation (BE).**

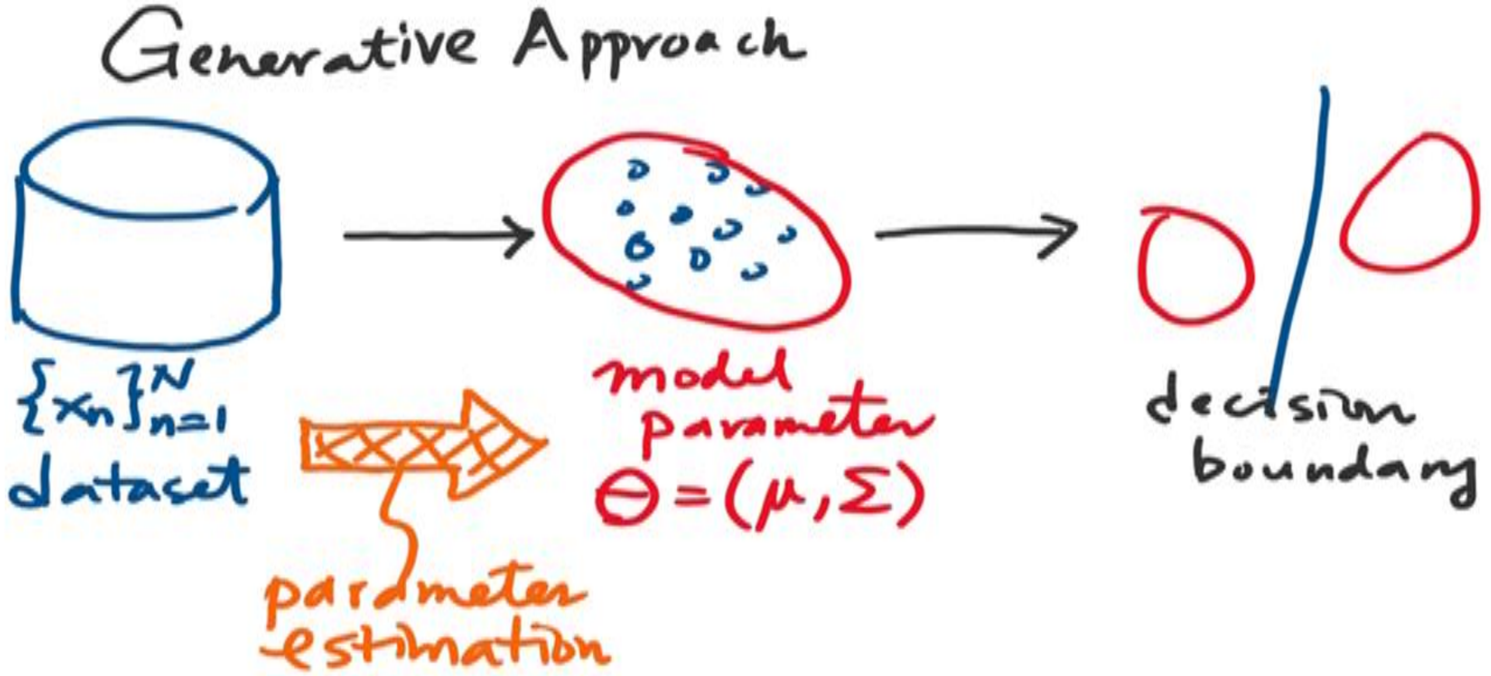
- Maximum likelihood and several other methods view the parameters as quantities whose values are fixed but unknown.

- Bayesian methods view the parameters as random variables having some known a priori distribution.

Introduction



Introduction



- The goal of parameter estimation is to determine $\boldsymbol{\Theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ from dataset
- This is *the step* where you use data

Source:ECE595 / STAT598: Machine Learning I Spring 2020 ,Stanley Chan

Introduction

ML Parameter Estimation

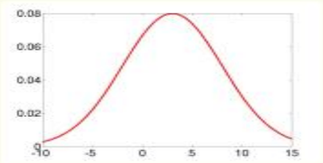
- Shape of probability distribution is known
  - Happens sometimes
- Labeled training data

- Need to estimate parameters of probability distribution from the training data

a lot is known "easier"

Example

respected fish expert says salmon's length has distribution $N(\mu_1, \sigma_1^2)$ and sea bass's length has distribution $N(\mu_2, \sigma_2^2)$



- Need to estimate parameters $\mu_1, \sigma_1^2, \mu_2, \sigma_2^2$
- Then design classifiers according to the bayesian decision theory

little is known "harder"

Source: CS434b/654b:Pattern Recognition Prof.Olga Veksler

Introduction

a simple example

p(Heads | Straight) =1/2

p(Tails | Straight) = 1/2

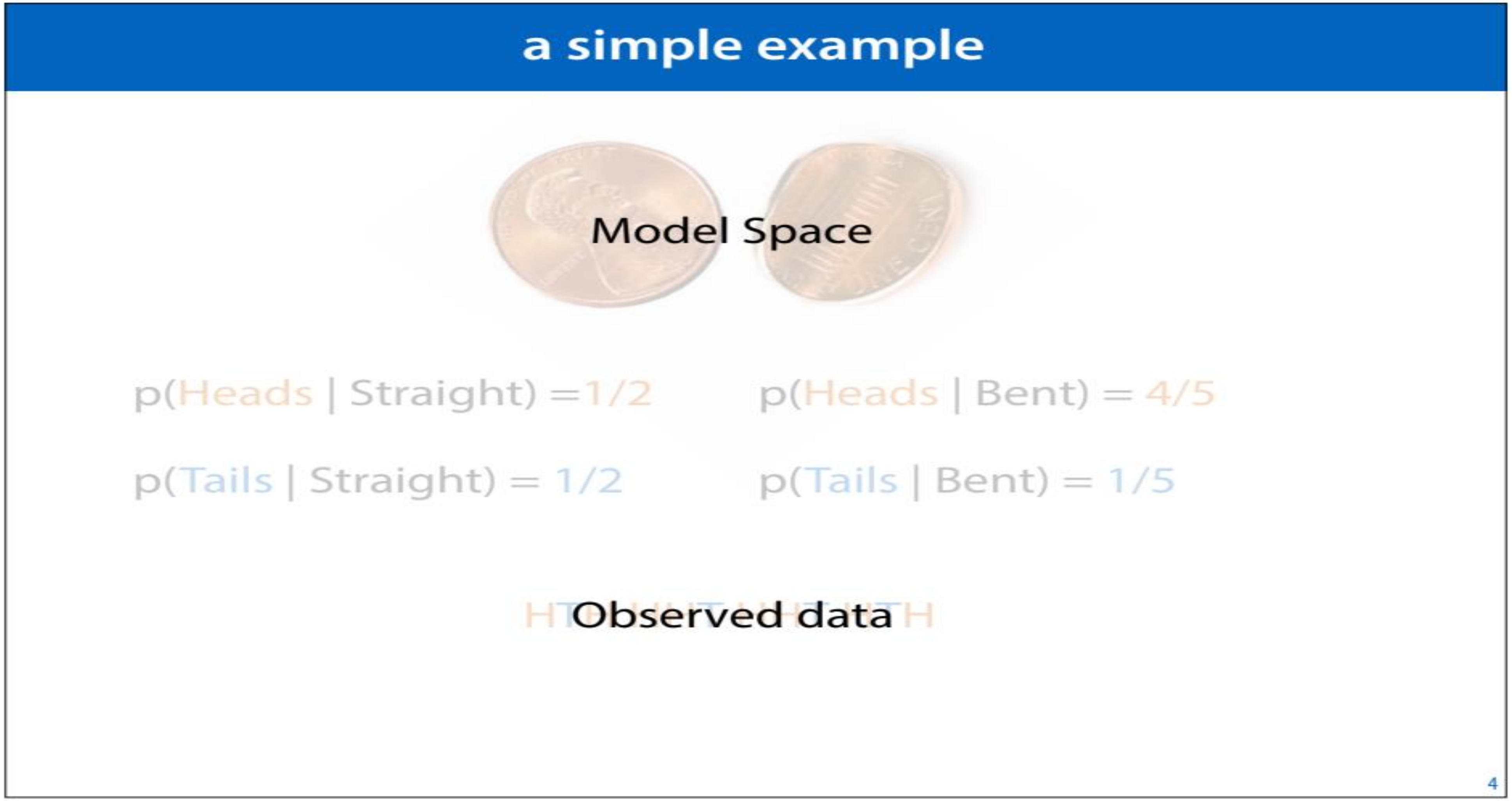
p(Heads | Bent) = 4/5

p(Tails | Bent) = 1/5

HTH HHT HHT HTH

3

Introduction



Introduction

maximum likelihood

$$\underset{\text{Coin}\in\{\text{Bent},\text{Straight}\}}{\arg\max}\; p(\text{HTHHHTHHTHTH} \mid \text{Coin})$$

$$\underset{\text{Model}\,\in\text{Model Space}}{\arg\max}\; p(\text{Data} \mid \text{Model}\;)$$

5

Introduction

which coin?

HTHHHTHHTHTH

$$p(D|\text{Bent}) = \frac{4}{5}\frac{1}{5}\frac{4}{5}\frac{4}{5}\frac{4}{5}\frac{1}{5}\frac{4}{5}\frac{4}{5}\frac{1}{5}\frac{4}{5}\frac{1}{5}\frac{4}{5} \approx 0.000268$$

$$p(D|\text{Straight}) = \frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2} \approx 0.000244$$

6

Parameter Estimation:

Main Methods

- **Maximum Likelihood** (ML)
  - Views the parameters θ as quantities whose values are fixed but unknown.
  - Estimates by maximizing the likelihood of obtaining the samples observed.

- **Bayesian Estimation** (BE)
  - Views the parameters θ as random variables having some known prior distribution $p(\theta)$.
  - Observing new samples D, converts the prior $p(\theta)$ to a posterior density $p(\theta/D)$ (i.e., the samples D revise our estimate over the parameters).

Maximum Likelihood Estimation(MLE)

- Suppose that we separate a collection of samples according to class so that we have c sets, $D_1, \ldots, D_c$.
- With the samples in D_j having been drawn independently according to the probability law $p(x|\omega_j)$.
- We say such samples are i.i.d. — independent identically distributed random variables.
- We assume that $p(x|\omega_j)$ has a known parametric form, and is therefore determined uniquely by the value of a parameter vector θ_j.

Maximum Likelihood Estimation

- For example, we might have $p(x|\omega_j) \sim N(\mu_j,\Sigma_j)$, where θj consists of the components of μj and Σj.

- To show the dependence of $p(x|\omega_j)$ on θj explicitly, we write $p(x|\omega_j)$ as $p(x|\omega_j, \theta_j)$.

- Our problem is to use the information provided by the training samples to obtain good estimates for the unknown parameter vectors θ_1, ..., θc associated with each category.

Maximum Likelihood Estimation

- Assume that, samples in Di give no information about θj

    - if i ≠ j that is, we shall assume that the parameters for the different classes are functionally independent.

- This permits us to work with each class separately.

- With this assumption we thus have c separate problems of the following form:

    - Use a set D of training samples drawn independently from the probability density $p(x|\theta)$ to estimate the unknown parameter vector θ.

Maximum Likelihood Estimation

Independence Across Classes

- We have training data for each class

- When estimating parameters for one class, will only use the data collected for that class
  - reasonable assumption that data from class c_i gives no information about distribution of class c_j

estimate parameters for distribution of salmon from	estimate parameters for distribution of bass from
	

Source: CS434b/654b:Pattern Recognition Prof.Olga Veksler

Maximum Likelihood Estimation

- Suppose that D contains n samples, x1, ...,xn. Then, since the samples were drawn independently, we have

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta}).$$

- A function of θ, p(D|θ) is called the likelihood of θ with respect to the set of samples.

- The maximum likelihood estimate of θ is, by definition, the value θˆ that maximizes p(D|θ).

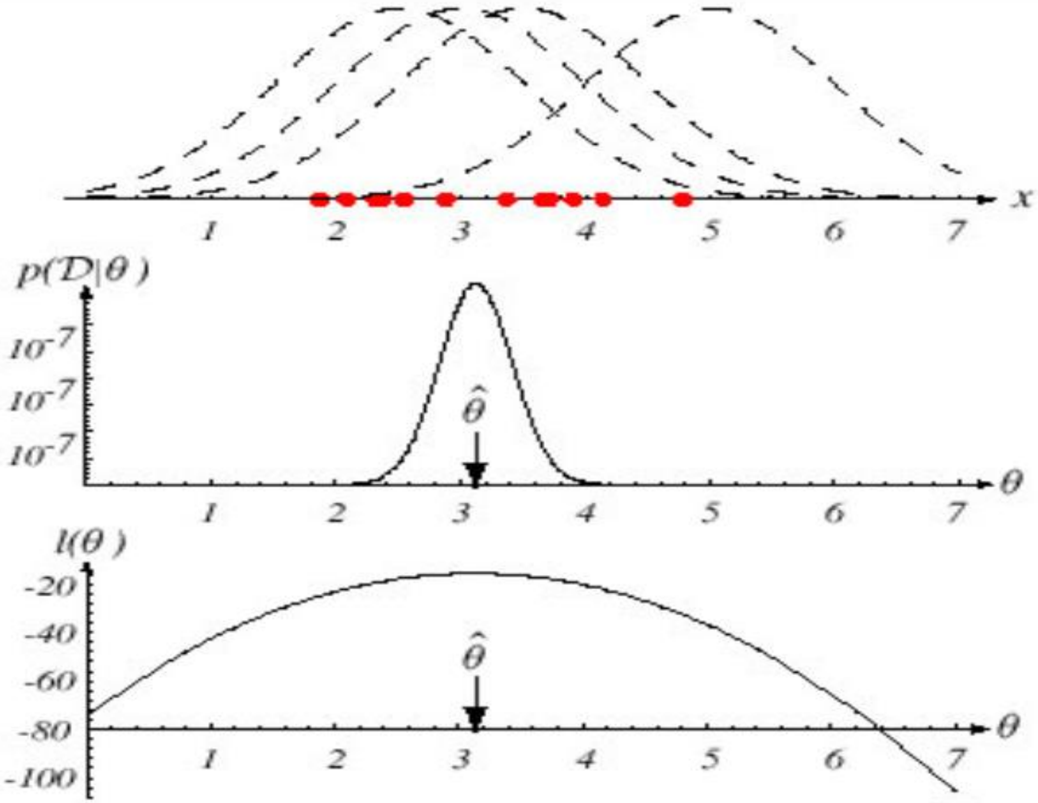


FIGURE 3.1. The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figure shows the likelihood $p(\mathcal{D}|\theta)$ as a function of the mean. If we had a very large number of training points, this likelihood would be very narrow. The value that maximizes the likelihood is marked $\hat{\theta}$; it also maximizes the logarithm of the likelihood—that is, the log-likelihood $l(\theta)$, shown at the bottom. Note that even though they look similar, the likelihood $p(\mathcal{D}|\theta)$ is shown as a function of θ whereas the conditional density $p(x|\theta)$ is shown as a function of x. Furthermore, as a function of θ, the likelihood $p(\mathcal{D}|\theta)$ is not a probability density function and its area has no significance. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.

2

Maximum Likelihood Estimation

Maximum Likelihood Parameter Estimation

- We have density $p(x)$ which is completely specified by parameters $\theta = [\theta_1, \ldots, \theta_k]$
  - If $p(x)$ is $N(\mu, \sigma^2)$ then $\theta = [\mu, \sigma^2]$
- To highlight that $p(x)$ depends on parameters θ we will write $p(x/\theta)$
  - Note overloaded notation, $p(x/\theta)$ is **not** a conditional density
- Let $D = \{x_1, x_2, \ldots, x_n\}$ be the n independent training samples in our data
  - If $p(x)$ is $N(\mu, \sigma^2)$ then $x_1, x_2, \ldots, x_n$ are iid samples from $N(\mu, \sigma^2)$

Source: CS434b/654b:Pattern Recognition Prof.Olga Veksler

Maximum Likelihood Estimation

Maximum Likelihood Parameter Estimation

- Consider the following function, which is called likelihood of θ with respect to the set of samples D

$$p(D\,|\,\theta) = \prod_{k=1}^{k=n} p(x_k\,|\,\theta) = F(\theta)$$

- Note if D is fixed $p(D/\theta)$ is **not** a density
- Maximum likelihood estimate (abbreviated MLE) of θ is the value of θ that maximizes the likelihood function $p(D/\theta)$

$$\boxed{\hat{\theta} = \arg\max_{\theta}(p(D\,|\,\theta))}$$

Source: CS434b/654b:Pattern Recognition Prof.Olga Veksler

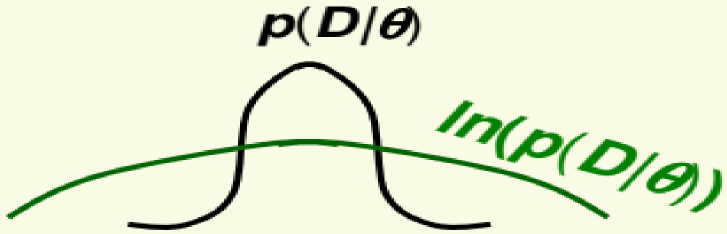
Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE)

- Instead of maximizing $p(D/\theta)$, it is usually easier to maximize $ln(p(D/\theta))$

- Since log is monotonic

$$\hat{\theta} = \arg\max_{\theta}(p(D\,|\,\theta)) =$$
$$= \arg\max_{\theta}(\ln p(D\,|\,\theta))$$



- To simplify notation, $ln(p(D/\theta))=l(\theta)$

$$\hat{\theta} = \arg\max_{\theta} l(\theta) = \arg\max_{\theta}\left(\ln\prod_{k=1}^{k=n} p(x_k\,/\,\theta)\right) = \arg\max_{\theta}\left(\sum_{k=1}^{n}\ln p(x_k\,/\,\theta)\right)$$

Source: CS434b/654b:Pattern Recognition Prof.Olga Veksler

Maximum Likelihood Estimation

MLE: Maximization Methods

- Maximizing $l(\theta)$ can be solved using standard methods from Calculus
- Let $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^t$ and let ∇_θ be the gradient operator

$$\nabla_\theta = \left[\frac{\partial}{\partial\theta_1}, \frac{\partial}{\partial\theta_2}, \ldots, \frac{\partial}{\partial\theta_p}\right]^t$$

- Set of necessary conditions for an optimum is:

$$\nabla_\theta l = 0$$

- Also have to check that θ that satisfies the above condition is maximum, not minimum or saddle point. Also check the boundary of range of θ

Source: CS434b/654b:Pattern Recognition Prof.Olga Veksler

The Gaussian Case: Unknown μ

MLE Example: Gaussian with unknown μ

- Fortunately for us, most of the ML estimates of any densities we would care about have been computed
- Let's go through an example anyway
- Let $p(x/\mu)$ be $N(\mu, \sigma^2)$ that is σ^2 is known, but μ is unknown and needs to be estimated, so $\theta = \mu$

$$\hat{\mu} = \arg\max_{\mu} l(\mu) = \arg\max_{\mu}\left(\sum_{k=1}^{n} \ln p(x_k / \mu)\right) =$$

$$= \arg\max_{\mu}\left(\sum_{k=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x_k-\mu)^2}{2\sigma^2}\right)\right)\right) =$$

$$= \arg\max_{\mu} \sum_{k=1}^{n}\left(-\ln\sqrt{2\pi\sigma} - \frac{(x_k-\mu)^2}{2\sigma^2}\right)$$

Source: CS434b/654b:Pattern Recognition Prof.Olga Veksler

The Gaussian Case: Unknown μ

MLE Example: Gaussian with unknown μ

$$\arg\max_{\mu}(l(\mu)) = \arg\max_{\mu}\sum_{k=1}^{n}\left(-\ln\sqrt{2\pi\sigma}-\frac{(x_k-\mu)^2}{2\sigma^2}\right)$$

$$\frac{d}{d\mu}(l(\mu)) = \sum_{k=1}^{n}\frac{1}{\sigma^2}(x_k-\mu) = 0 \;\Rightarrow\; \sum_{k=1}^{n}x_k - n\mu = 0 \;\Rightarrow$$

$$\Rightarrow \quad \hat{\mu} = \frac{1}{n}\sum_{k=1}^{n}x_k$$

- Thus the ML estimate of the mean is just the average value of the training data, very intuitive!
  - average of the training data would be our guess for the mean even if we didn't know about ML estimates

Source: CS434b/654b:Pattern Recognition Prof.Olga Veksler

The Gaussian Case: Unknown μ

MLE Example: Gaussian with unknown μ

$$\arg\max_{\mu}(l(\mu)) = \arg\max_{\mu}\sum_{k=1}^{n}\left(-\ln\sqrt{2\pi\sigma} - \frac{(x_k-\mu)^2}{2\sigma^2}\right)$$

$$\frac{d}{d\mu}(l(\mu)) = \sum_{k=1}^{n}\frac{1}{\sigma^2}(x_k-\mu) = 0 \Rightarrow \sum_{k=1}^{n}x_k - n\mu = 0 \Rightarrow$$

$$\Rightarrow \hat{\mu} = \frac{1}{n}\sum_{k=1}^{n}x_k$$

- Thus the ML estimate of the mean is just the average value of the training data, very intuitive!
  - average of the training data would be our guess for the mean even if we didn't know about ML estimates

Source: CS434b/654b:Pattern Recognition Prof.Olga Veksler

The Gaussian Case: Unknown μ and Σ

MLE for Gaussian with unknown μ, σ^2

- Similarly it can be shown that if $p(x/\mu,\sigma^2)$ is $N(\mu, \sigma^2)$, that is x both mean and variance are unknown, then again very intuitive result

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})^2$$

- Similarly it can be shown that if $p(x/\mu,\Sigma)$ is $N(\mu, \Sigma)$, that is x is a multivariate gaussian with both mean and covariance matrix unknown, then

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k \qquad \hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})(x_k - \hat{\mu})^t$$

Source: CS434b/654b:Pattern Recognition Prof.Olga Veksler

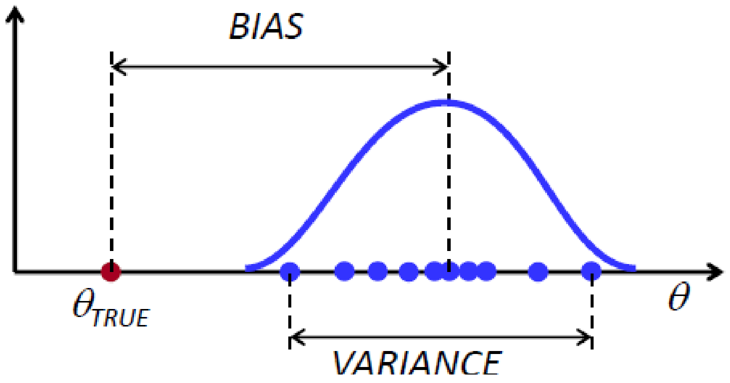
Biased and Unbiased Estimators

- An estimator of a parameter is biased if the expected value of the estimate is different from the true value of the parameters.

- An estimator of a parameter is unbiased if the expected value of the estimate is the same as the true value of the parameters.

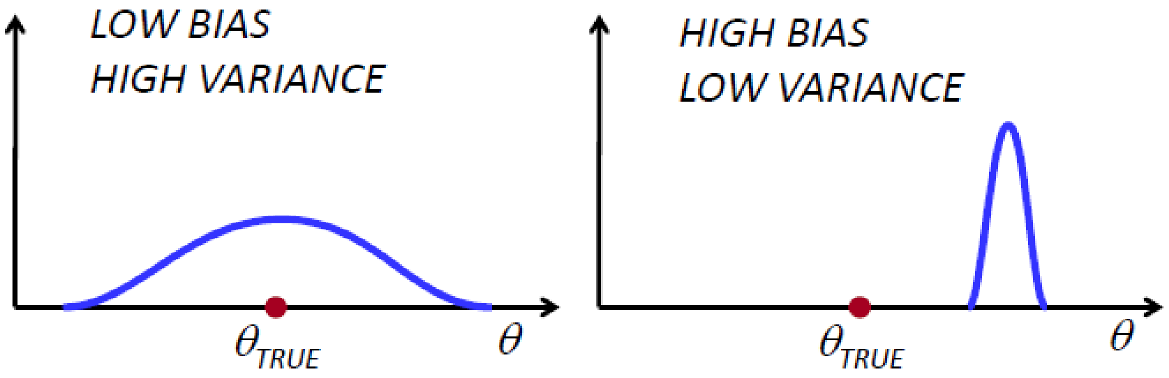
Bias and Variance

- **How good are the ML estimates?**
  - Two measures of "goodness" are used for statistical estimates
  - **Bias**: how close is the estimate to the true value?
  - **Variance**: how much does it change for different datasets?



Bias and Variance

- The bias-variance tradeoff: in most cases, you can only decrease one of them at the expense of the other



Biased and Unbiased Estimates

- An estimate $\hat{\boldsymbol{\theta}}$ is unbiased when

$$E[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$$

- The ML estimate $\hat{\boldsymbol{\mu}}$ is unbiased, i.e.,

$$E[\hat{\boldsymbol{\mu}}] = \boldsymbol{\mu}$$

- The ML estimates $\hat{\boldsymbol{\sigma}}$ and $\hat{\Sigma}$ are biased:

$$E[\hat{\boldsymbol{\sigma}}^2] = \frac{n-1}{n}\sigma^2 \qquad E[\hat{\Sigma}] = \frac{n-1}{n}\Sigma$$

Gaussian Case: Unknown Mean

- **Consider the case where only the mean, θ = μ, is unknown:**

$$\sum_{k=1}^{n} \nabla_{\boldsymbol{\theta}} \ln\left(p\left(\mathbf{x}_k \middle| \boldsymbol{\theta}\right)\right) = 0$$

$$\ln(p(\mathbf{x}_\mathrm{k}|\boldsymbol{\theta})) = \ln[\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp[\frac{-1}{2}(\mathbf{x}_k - \boldsymbol{\theta})^{\mathrm{t}} \Sigma^{-1}(\mathbf{x}_k - \boldsymbol{\theta})]$$

$$= -\frac{1}{2}\ln[(2\pi)^d|\Sigma|] - \frac{1}{2}(\mathbf{x}_k - \boldsymbol{\theta})^{\mathrm{t}} \Sigma^{-1}(\mathbf{x}_k - \boldsymbol{\theta})$$

which implies: $\nabla_{\boldsymbol{\theta\mu}} \ln(p(\mathbf{x}_\mathrm{k}|\boldsymbol{\theta})) = \Sigma^{-1}(\mathbf{x}_k - \boldsymbol{\theta})$

because:

$$\frac{\partial}{\partial \boldsymbol{\theta}}\left\{[-\frac{1}{2}\ln[(2\pi)^d|\Sigma|] - \frac{1}{2}(\mathbf{x}_k - \boldsymbol{\theta})^{\mathrm{t}} \Sigma^{-1}(\mathbf{x}_k - \boldsymbol{\theta})]\right\}$$

$$= \frac{\partial}{\partial \boldsymbol{\theta}}[-\frac{1}{2}\ln[(2\pi)^d|\Sigma|] - \frac{\partial}{\partial \boldsymbol{\theta}}[\frac{1}{2}(\mathbf{x}_k - \boldsymbol{\theta})^{\mathrm{t}} \Sigma^{-1}(\mathbf{x}_k - \boldsymbol{\theta})]$$

$$= \Sigma^{-1}(\mathbf{x}_k - \boldsymbol{\theta})$$

Gaussian Case: Unknown Mean

- **Substituting into the expression for the total likelihood:**

$$\nabla_{\boldsymbol{\theta}} l = \sum_{k=1}^{n} \nabla_{\boldsymbol{\theta}} \ln\left(p\left(\mathbf{x}_k \middle| \boldsymbol{\theta}\right)\right) = \sum_{k=1}^{n} \Sigma^{-1}(\mathbf{x}_k - \boldsymbol{\theta}) = 0$$

- **Rearranging terms:** $\sum_{k=1}^{n} \Sigma^{-1}(\mathbf{x}_k - \hat{\boldsymbol{\theta}}) = 0$

$$\sum_{k=1}^{n} (\mathbf{x}_k - \hat{\boldsymbol{\theta}}) = 0$$

$$\sum_{k=1}^{n} \mathbf{x}_k - \sum_{k=1}^{n} \hat{\boldsymbol{\theta}} = 0$$

$$\sum_{k=1}^{n} \mathbf{x}_k - n\hat{\boldsymbol{\theta}} = 0$$

$$\hat{\boldsymbol{\theta}} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k$$

Gaussian Case: Unknown Mean and Variance

- Let $\boldsymbol{\theta} = [\mu,\sigma^2]$. The log likelihood of a SINGLE point is:

$$\ln(p(x_k|\boldsymbol{\theta})) = -\frac{1}{2}\ln[(2\pi)\theta_2] - \frac{1}{2}(x_k - \theta_1)^t\theta_2^{-1}(x_k - \theta_1)$$

$$\nabla_{\boldsymbol{\theta}} l = \nabla_{\boldsymbol{\theta}} \ln(p(x_k|\boldsymbol{\theta})) = \begin{bmatrix} \dfrac{1}{\theta_2}(x_k - \theta_1) \\ -\dfrac{1}{2\theta_2} + \dfrac{(x_k-\theta_1)^2}{2\theta_2^2} \end{bmatrix}$$

- The full likelihood leads to:

$$\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0$$

$$\sum_{k=1}^{n} -\frac{1}{2\hat{\theta}_2} + \frac{(x_k-\hat{\theta}_1)^2}{2\hat{\theta}_2^2} = 0 \Rightarrow \sum_{k=1}^{n}(x_k-\hat{\theta}_1)^2 = \sum_{k=1}^{n}\hat{\theta}_2$$

Gaussian Case: Unknown Mean and Variance

- **This leads to these equations:** $\hat{\theta}_1 = \hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k$

$$\hat{\theta}_2 = \hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})^2$$

- **In the multivariate case:** $\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} \mathrm{x}_k$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(\mathrm{x}_k - \hat{\mu})(\mathrm{x}_k - \hat{\mu})^{\mathrm{t}}$$

- **The true covariance is the expected value of the matrix** $(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t$ **, which is a familiar result.**

Convergence of the Mean

- **Does the maximum likelihood estimate of the variance converge to the true value of the variance? Let's start with a few simple results we will need later.**

- **Expected value of the ML estimate of the mean:**

$$
\begin{aligned}
E[\hat{\mu}] &= E[\frac{1}{n}\sum_{i=1}^{n} x_i] \\
&= \frac{1}{n}\sum_{i=1}^{n} E[x_i] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu = \mu
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{var}[\hat{\mu}] &= E[\hat{\mu}^2] - (E[\hat{\mu}])^2 \\
&= E[\hat{\mu}^2] - \mu^2 \\
&= E[\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{j=1}^{n} x_j\right)] - \mu^2 \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} E[x_i x_j]\right)^2 - \mu^2
\end{aligned}
$$

Variance of the ML Estimate of the Mean

- The expected value of $x_i x_j$, $E[x_i x_j]$, will be μ^2 for $i \neq j$ and $\mu^2 + \sigma^2$ otherwise since the two random variables are independent.

- The expected value of x_i^2 will be $\mu^2 + \sigma^2$.

- Hence, in the summation above, we have n^2-n terms with expected value μ^2 and n terms with expected value $\mu^2 + \sigma^2$.

- Thus,

$$\mathrm{var}[\hat{\mu}] = \frac{1}{n^2}\left(\left(n^2 - n\right)\mu^2 + n\left(\mu^2 + \sigma^2\right)\right) - \mu^2 = \frac{\sigma^2}{n}$$

which implies:

$$E[\hat{\mu}^2] = \mathrm{var}[\hat{\mu}] + (E[\hat{\mu}])^2 = \frac{\sigma^2}{n} + \mu^2$$

- We see that the variance of the estimate goes to zero as n goes to infinity, and our estimate converges to the true estimate (error goes to zero).

Variance Relationships

- **We will need one more result:**

$$\sigma^2 = E[(x-\mu)^2 = E[x^2] - 2E[x]\mu + E[\mu^2]$$
$$= E[x^2] - 2\mu^2 + E[\mu^2]$$
$$= E[x^2] - \mu^2$$
$$= \sum_{i=1}^{n} x_i^2 - (\frac{1}{n}\sum_{i=1}^{n} x_i)^2$$

Note that this implies:

$$\sum_{i=1}^{n} x_i^2 = \sigma^2 + \mu^2$$

- **Now we can combine these results. Recall our expression for the ML estimate of the variance:**

$$\hat{\sigma}^2 = E[\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2]$$

Covariance Expansion

- **Expand the covariance and simplify:**

$$\hat{\sigma}^2 = E[\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 = \frac{1}{n}E[\sum_{i=1}^{n}(x_i^2 - 2x_i\hat{\mu} + \hat{\mu}^2)]$$

$$= \frac{1}{n}\sum_{i=1}^{n}(E[x_i^2] - 2E[x_i\hat{\mu}] + E[\hat{\mu}^2])$$

$$= \frac{1}{n}\sum_{i=1}^{n}((\sigma^2 + \mu^2) - 2E[x_i\hat{\mu}] + (\mu^2 + \sigma^2/n))$$

- **One more intermediate term to derive:**

$$E[x_i\hat{\mu}] = E[x_i \sum_{j=1}^{n} x_j] = \frac{1}{n}\sum_{j=1}^{n} E[x_i x_j] = \frac{1}{n}(\sum_{\substack{j=1 \\ i\neq j}}^{n} E[x_i x_j] + E[x_i x_i])$$

$$= \frac{1}{n}((n-1)\mu^2 + \left(\mu^2 + \sigma^2\right)) = \frac{1}{n}((n\mu^2 + \sigma^2) = \mu^2 + \frac{\sigma^2}{n}$$

Biased Variance Estimate

- **Substitute our previously derived expression for the second term:**

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}((\sigma^2+\mu^2) - 2E[x_i\hat{\mu}] + (\mu^2 + \sigma^2/n))$$

$$= \frac{1}{n}\sum_{i=1}^{n}((\sigma^2+\mu^2) - 2(\mu^2 + \sigma^2/n) + (\mu^2 + \sigma^2/n))$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\sigma^2+\mu^2 - 2\mu^2 + \mu^2 - 2\sigma^2/n + \sigma^2/n)$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\sigma^2 - \sigma^2/n)$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\sigma^2 - \sigma^2/n) = \frac{1}{n}\sum_{i=1}^{n}\sigma^2(1-1/n) = \frac{1}{n}\sum_{i=1}^{n}\sigma^2\frac{(n-1)}{n}$$

$$= \frac{(n-1)}{n}\sigma^2$$

Expectation Simplification

- **Therefore, the ML estimate is biased:**

$$\hat{\sigma}^2 = E[\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2] = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

However, the ML estimate converges .

- **An unbiased estimator is:**

$$\mathbf{C} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^t$$

Page 1

I) Sample variance of ML estimator

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\therefore E(s^2) = E\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$$

$$= \frac{1}{n}E\left[\sum_{i=1}^{n} x_i^2 - 2x_i\bar{x} + \bar{x}^2\right]$$

$$= \frac{1}{n}E\left[\sum x_i^2 - 2\sum x_i\bar{x} + \sum \bar{x}^2\right]$$

$$= \frac{1}{n}E\left[\sum x_i^2 - 2(\bar{x}n)\bar{x} + n\bar{x}^2\right]$$

$$= \frac{1}{n}E\left[\sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right]$$

$$= \frac{1}{n}E\left[\sum x_i^2 - n\bar{x}^2\right]$$

$$= \frac{1}{n}\sum E(x^2) - E(\bar{x}^2)$$

$$= \frac{1}{n}\times nE(x^2) - E(\bar{x}^2)$$

$$\boxed{E(s^2) = E(x^2) - E(\bar{x}^2)}$$

Now, $\sigma_x^2 = E(x^2) - [E(x)]^2$

or $\sigma_x^2 = E(x^2) - \mu^2$

$$\therefore \boxed{E(x^2) = \sigma^2 + \mu^2}$$

Now, $\boxed{\sigma_{\bar{x}}^2 = E(\bar{x}^2) - [E(\bar{x})]^2}$

Now, $Var(x_1 + x_2 + x_3 + \cdots + x_n) = n\cdot\sigma_x^2$

$$Var\left(\frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}\right) = \frac{n\sigma_x^2}{n^2} = \frac{\sigma_x^2}{n}$$

$$= \sigma_{\bar{x}}^2$$

[P.T.O]

Page:2

$$\therefore \quad \sigma_{\bar{x}}^{2} = E(\bar{x}^2) - [E(\bar{x})]^2$$

$$\frac{\sigma_x^2}{n} = E(\bar{x}^2) - \underbrace{[E(\bar{x})]^2}_{[E(x)]^2}$$

$$\therefore \quad \frac{\sigma_x^2}{n} = E(\bar{x}^2) - [E(x)]^2$$

Here $E(x)$ is μ

$$\therefore \quad \frac{\sigma_x^2}{n} = E(\bar{x}^2) - \mu^2$$

$$\Rightarrow E(\bar{x}^2) = \frac{\sigma_x^2}{n} + \mu^2$$

Now, we know

$$E(s^2) = E(x^2) - E(\bar{x}^2)$$

$$= (\sigma^2 + \mu^2) - \left(\frac{\sigma^2}{n} + \mu^2\right)$$

$$= \frac{n\sigma^2 + n\mu^2 - \sigma^2 + n\mu^2}{n}$$

$$= \frac{(n-1)\sigma^2}{n}$$

$$\therefore \quad E(s^2) = \left(\frac{n-1}{n}\right)\sigma^2$$