Bayesian Learning

- Machine Learning by Mitchell-Chp. 6
- Ethem Chp. 3 (Skip 3.6)
- Pattern Recognition & Machine Learning by Bishop Chp. 1

(Pics mostly from the Bishop book)

- Berrin Yanikoglu

- last edited March 2025

1. **We have used binary/categorical attributes in the PlayTennis example, what about real-valued attributes?**
   - E.g if temperature was a continuous variable
   - We will see that we can use Naïve Bayes with continuous attributes or a mix of attribute types

2. **How do we estimate the class conditional probabilities $P(X_i \mid C_j)$ in general general.**
   - We have estimated P(Outlook = Sunny | Play =Yes) by counting the number of Sunny days and dividing by all days in Play=Yes class.
   - We will learn about **Maximum Likelihood Estimation approach**

Slides thanks to Oznur Tastan

Naïve Bayes Classifier with a Single Continuous Attribute

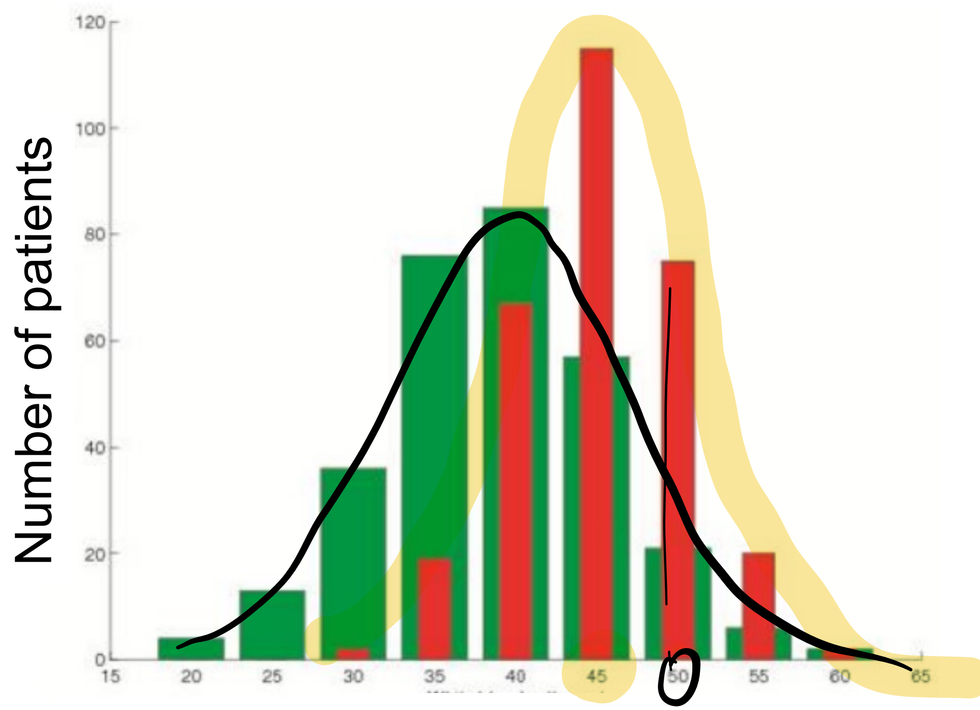
Bayes Classifier

- Aim to classify if a patient has diabetes, classify into one of two classes (yes Y = 1; no Y = 0)
- Conduct several tests on the patient and got X
- Given patient's result X, compute the posteror probability using Bayes Rule

$$\underbrace{\mathbf{P}(Y \mid X)}_{\text{Posterior}} = \frac{\overbrace{\mathbf{P}(X \mid Y)}^{\text{Class likelihood}}\overbrace{\mathbf{P}(Y)}^{\text{Class priors}}}{\underbrace{\mathbf{P}(X)}_{\text{Evidence}}}$$

Classification: Diabetes Example

- Let's start with the simplest case where the input is 1-dimensional, only one feature
- We need to choose a probability distribution for the P(X | Y)

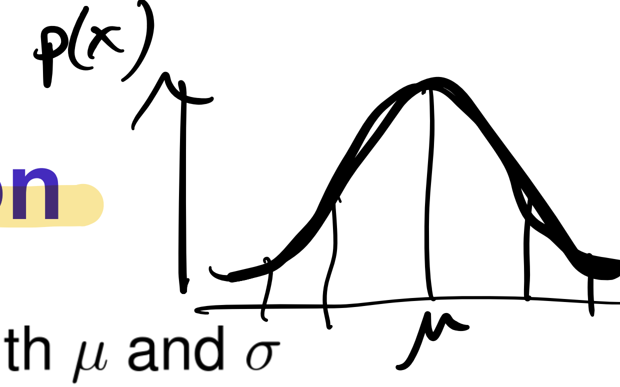


Red: diabetes
Green: normal

$x = 50$

Slide adated from R. Urtasun

Gaussian(Normal) Distribution

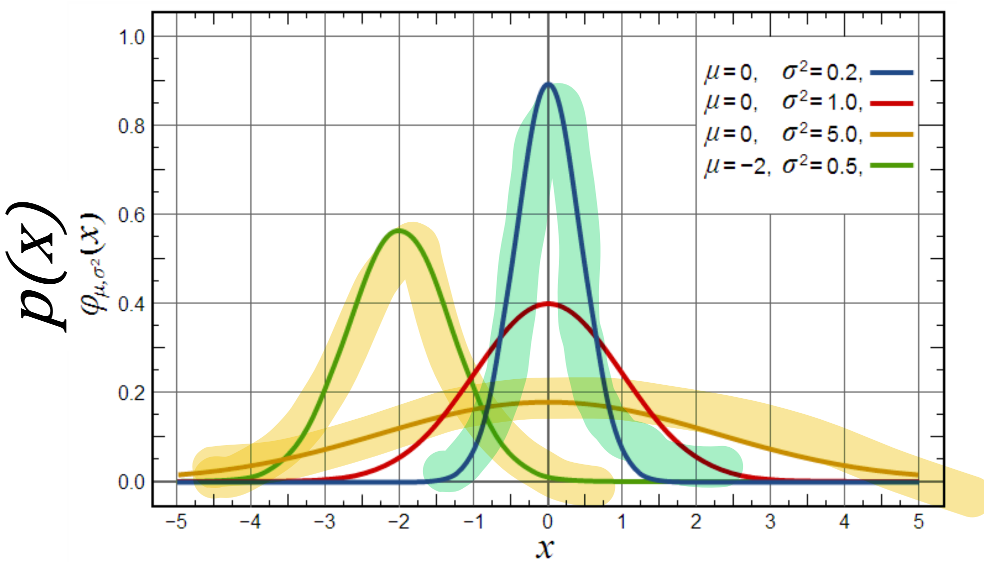


- The probability density function parameterized with μ and σ
  $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$

- The probability that X will all into the interval (a, b) is given by
  $\int_a^b p(x)dx$

- Expected, or mean value of X, $E[X]$
  $E[X] = \mu$

- Variance of X is
  $Var(X) = \sigma^2$



Gausian Bayes Classifier

- Assume the white blood cell count is **normally distributed**.

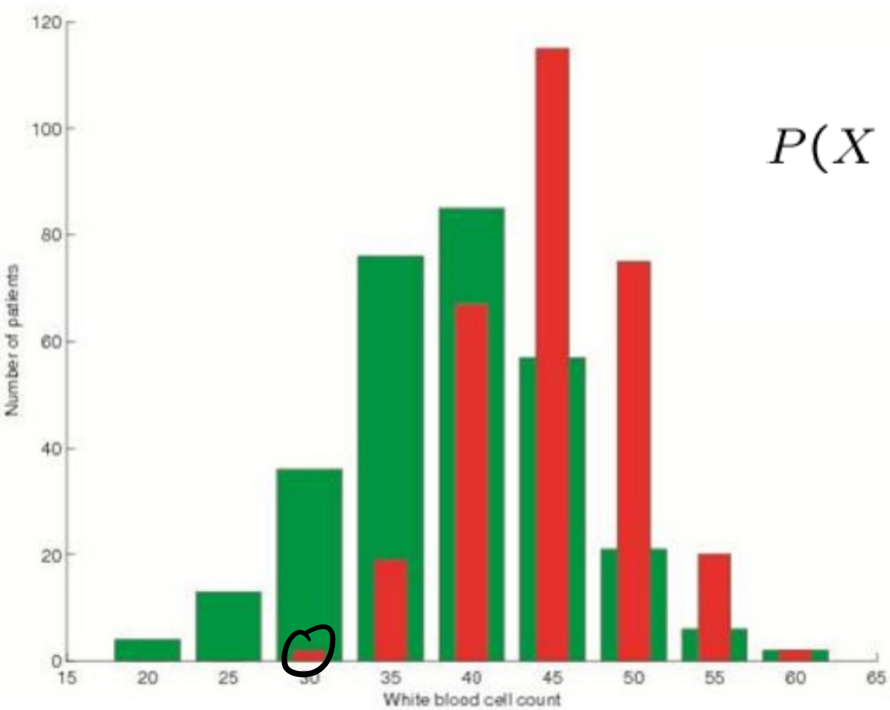
$$P(X = x|Y = y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x-\mu_y)^2}{2\sigma_y^2}\right)$$

$$e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

  - Notice we assume a **different Gaussian for each class**
  - These are **class conditional** Gausians

Fitting Gausians to the Data

- How can I fit a Gaussian distribution to my data?



$$P(X = x|Y = y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x-\mu_y)^2}{2\sigma_y^2}\right)$$

$\mu_{diabetes} = ?$

$\sigma_{diabetes} = ?$

x 30 diabetes (red)
x 35 " "
x 35 " "
x 35

x 35 healthy (+)

MLE for Gausians

- We will estimate the parameters of a Gaussian distribution using the **Maximum Likelihood Estimation** (in later slides):

MLE estimates of parameters for a Gaussian distribution:

$$\mu_{MLE} = \frac{1}{N}\sum_{n=1}^{N} x^{(n)}$$

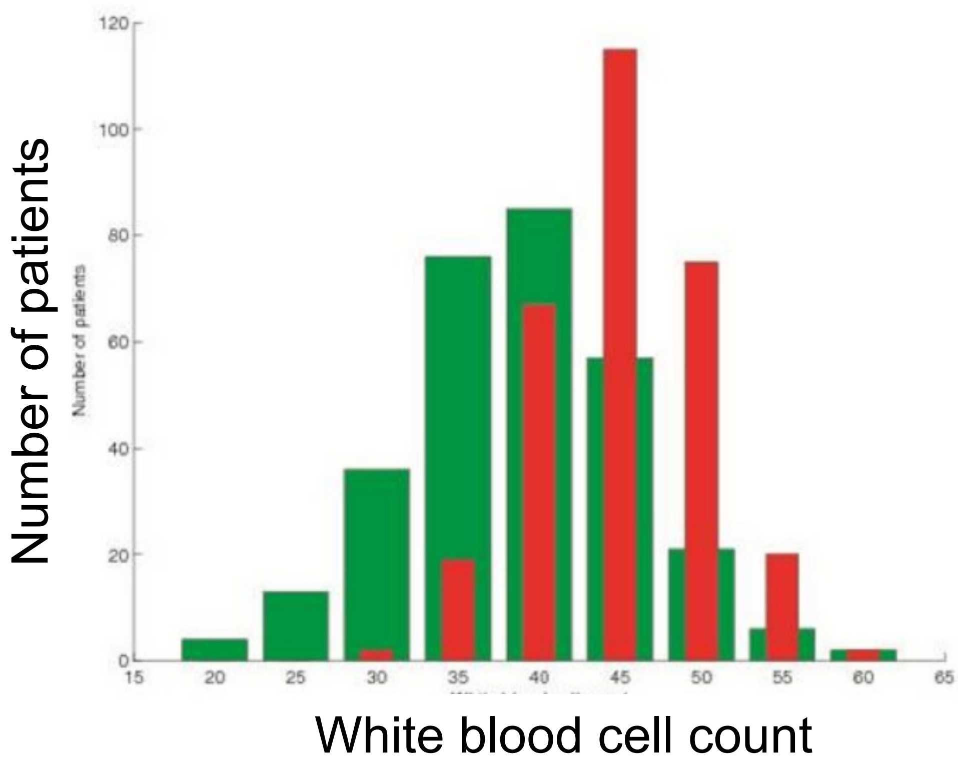
$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N} (x^{(n)} - \mu)^2$$

sum all $x^{(n)}$'s
// & divide by N

N samples in
red class

Diabetes Example – NB Classification

Slide adated from R. Urtasun

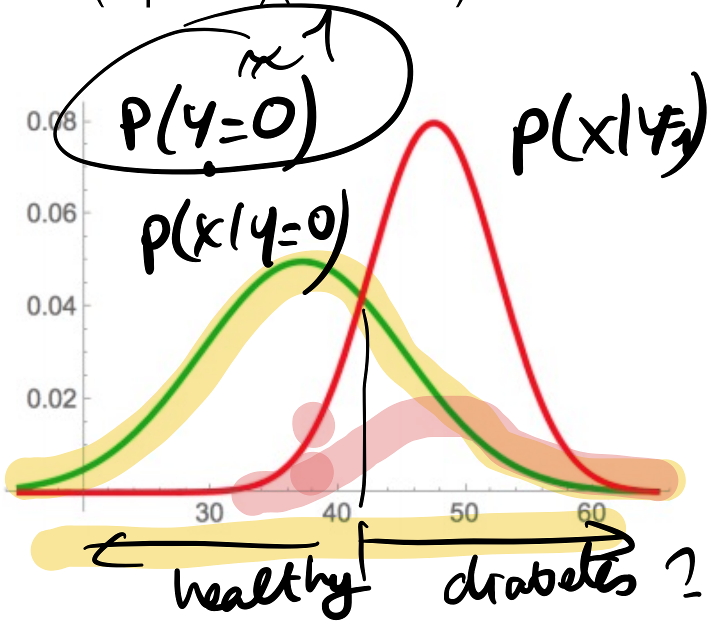


- Doctor has a prior p(Y = 0) = 0.8
- A new patient comes in, the doctor measures X= 48
- Does the patient have diabetes?

Bayes Classifier

P(X | Y =0) (no diabetes)
P(X | Y =1) (diabetes)

$$P(X=x|Y=y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x-\mu_y)^2}{2\sigma_y^2}\right)$$



p(x|y=1) · p(y=1)

- Compute P(X= 48 | Y =0) and P (X=48| Y= 1) via the estimated **class conditional Gaussian distributions**
- Compute posteriors P (Y = 0 | X = 48) and P (Y = 1 | X = 48) via **Bayes rule**.
- Choose the class with maximum posterior

1a) $\nabla = \begin{bmatrix} 2x-5, & 2y \end{bmatrix}$

$\nabla \|_{(1,1)} = \begin{bmatrix} -3 & 2 \end{bmatrix}$

$P_1 = \underline{P0} - 0.1$

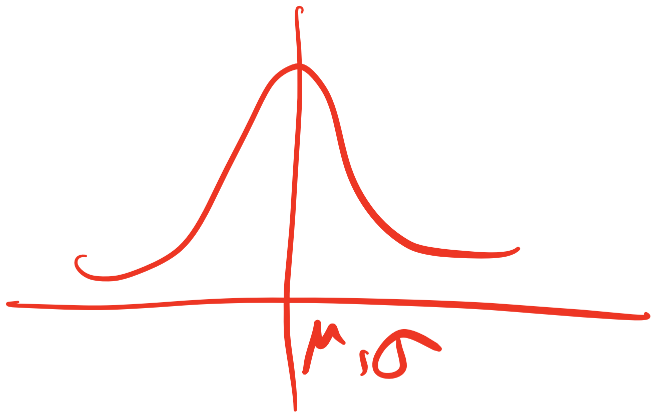
$\nabla_{(1,1)}$

b) $\nabla = \begin{bmatrix} \phi \end{bmatrix} \rightarrow$

$P(C_i|x)$ — NBayes —

Discrete X :

Cts x :



$$\mathcal{N}(x|\mu_i, \sigma_i) = p(x|C_i)$$

$$P(C_i) \cdot p(x|C_i)$$

Cts x_is

$$\prod_{d=1}^{k} p(x_d|C_i)$$

Naïve Bayes Classifier with Multiple Continuous Attributes

$P(X \mid C_i)$ $\quad C_i =$ $P(\cdot)$ $P(\cdot)$

Multiple Continuous Features

Understanding cognitive function from images of neuronal activity (real number) from 20,000 locations in the brain.

- Is the person reading a sentence or viewing a picture?
- Reading the word "Knife" or "Apartment"?
- Viewing a vertical or horizontal line?

$P(Yes)$ $\quad P(No)$

$P(p_1 = 230, p_2 = 50 \mid P) \Rightarrow P(p_1 = 230 \mid P) \times P(p_2 = 50 \mid P)$

$= X = \begin{bmatrix} p_1 \\ \vdots \\ p_{20{,}000} \end{bmatrix}$

Related Video (experiment done at CMU by T. Mitchell)

X_i s are Continuous

- Naïve Bayes model:

We just need to model this:

$$P(Y=y|X_1,\ldots,X_n) = \frac{P(Y=y)\prod_i P(X_i|Y=y)}{\sum_j P(Y=y_j)\prod_i P(X_i|Y=y_j)}$$

X_i s are Continuous

- We can still use the Naïve Bayes model

We just need to model this

$$P(Y=y|X_1,\ldots,X_n) = \frac{P(Y=y)\prod_i P(X_i|Y=y)}{\sum_j P(Y=y_j)\prod_i P(X_i|Y=y_j)}$$

- Common approach: assume $P(X_i|Y=y_k)$ follows a Gaussian (Normal) distribution

$$\mathbf{P}\left(X_i = x \mid Y = y_k\right) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} exp^{-\frac{1}{2}\left(\frac{x-\mu_{ik}}{\sigma_{ik}}\right)^2}$$

Gaussian Naïve Bayes

- Gausian Naive Bayes assume:

$$\mathbf{P}\left(X_i = x \,\middle|\, Y = y_k\right) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} exp^{-\frac{1}{2}\left(\frac{x-\mu_{ik}}{\sigma_{ik}}\right)^2}$$

Gaussian Naïve Bayes

- Train Naïve Bayes (examples)

  for each value y_k

  estimate $\pi_k \equiv P(Y = y_k)$

  for each attribute X_i estimate $P(X_i | Y = y_k)$

  - conditional mean μ_{ik} , variance σ_{ik}

- Classify (X^{new})

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$Y^{new} \leftarrow \arg\max_{y_k} \; \pi_k \prod_i \mathcal{N}(X_i^{new}; \mu_{ik}, \sigma_{ik})$$

$X \sim \text{Bernoulli}(p)$

$p: 0.8$ for Heads

$p: 0.5$ " "

HHHHTHHHHH

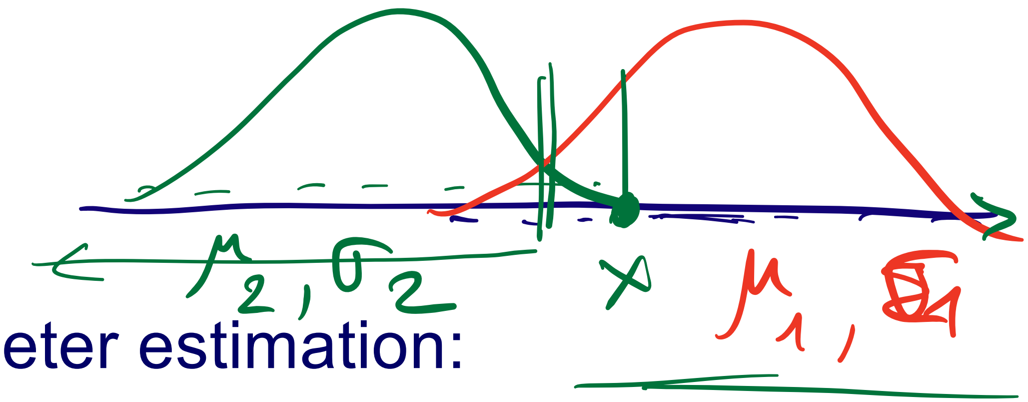
Find p that maximizes the likelihood of the observed data.

$$\text{Likelihood}(p) = \prod p^{H} \cdot (1-p)^{T}$$

H: 1

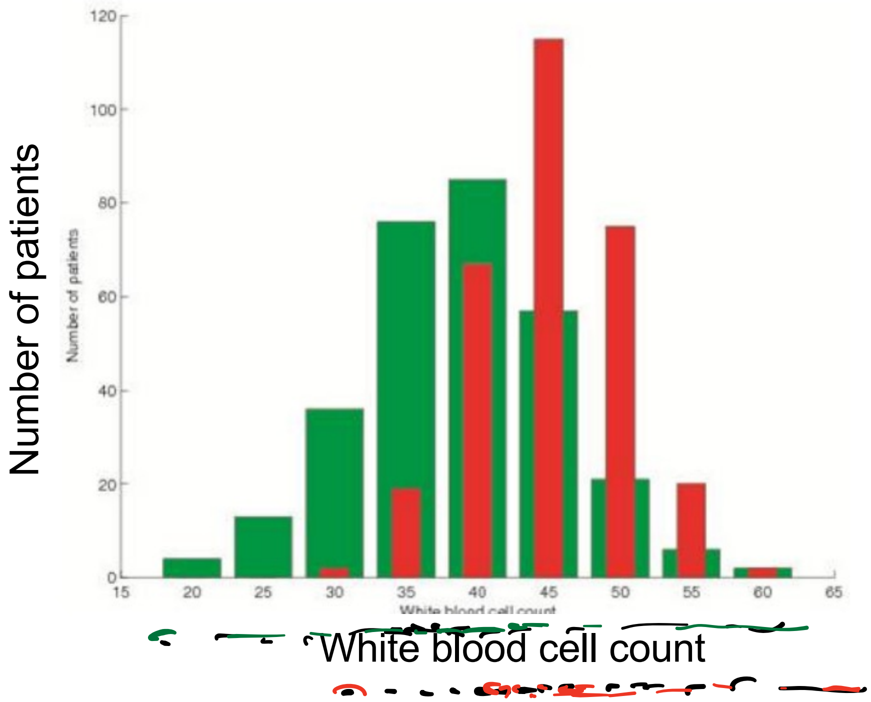
T: 0

Maximum Likelihood Estimate for 1D-Gaussian distributions



- Two approaches in parameter estimation:
  - Maximum Likelihood Estimation (MLE)
  - Maximum a Posteriori (MAP) estimate

healthy



Intuition: A mean of 90 when the red data is centered around 45 is possible, but unlikely.

Task is find μ_1, σ_1
μ_2, σ_2
estimates.

$\rightarrow$ MLE

Maximum Likelihood

40

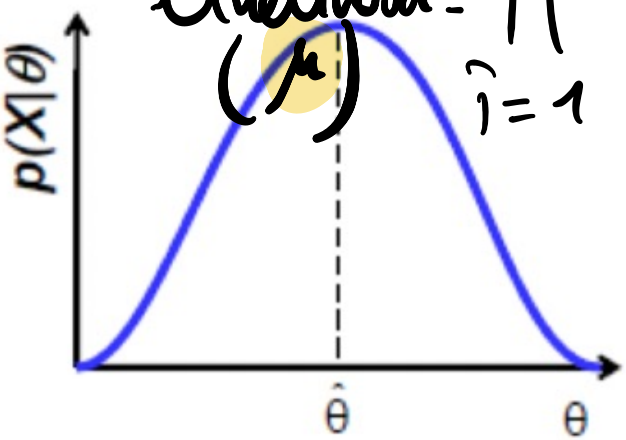
- The Maximum Likelihood (ML) solution seeks the solution that best explains the dataset X

$\mu = 110$

$\mu = 40$

$$\underset{\mu}{\hat{\theta}} = \underset{\theta}{\text{argmax}}\underbrace{[p(X\,|\,\theta)]}_{\text{likelihood}}$$

$$\text{likelihood}(\mu) = \prod_{i=1}^{N}\left(\frac{1}{\sqrt{2\pi\sigma}}\cdot e^{-\left(\frac{x^{(i)}-\mu}{\sigma}\right)^2}\right)$$



$p(X \text{ or } x \,|\, \theta)$

$p(D\,|\,\theta)$: function of θ.

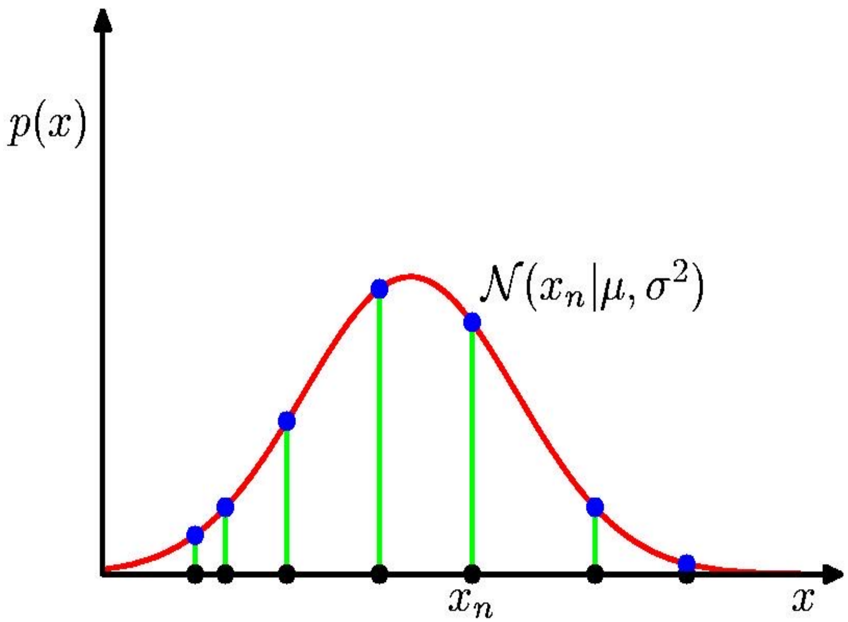
likelihood -

$p(X = x_0)$: density

Maximum Likelihood Estimation for 1D Gaussian Distribution

Given N data points **x**$=\{x_1, x_2, ..., x_N\}$, where x_i is assumed to be normally distributed, the Maximum Likelihood Estimation tries to
find the parameters μ and σ that maximizes
the probability of seeing the observed data with the given parameters (likelihood).

- ☐ need to write the probability of the data for assumed μ and σ
- ☐ find the values of μ and σ maximizing this probability.



all data

$$p(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\sigma^2)$$

Assuming iid data

To maximize or minimize a function w.r.t (with respect to) some parameters,

- ☐ **we need to take the derivative of the function w.r.t the parameter and set to 0** (because the min or max of a function is a local minima/maxima)
- ☐ **solve for the unknown parameter.**

Here we maximize:

$$P(\cancel{x} \mid \mu, \sigma) = \prod_{n=1}^{N} p(x_n \mid \mu, \sigma^2) = \prod_{n=1}^{N} N(x_n \mid \mu, \sigma^2)$$

This is a topic called parameter estimation – Ethem Chapter 4 and other books. We will cover only the necessary parts in these slides.

- In fact, we will maximize log of that:

  Find parameters μ and σ that maximizes log $P(\mathbf{x} \mid \mu, \sigma)$

- If we plot the likelihood as a function of a parameter θ (e.g. μ or σ)

$$\hat{\theta} = \operatorname{argmax}\left[p(X \mid \theta)\right] = \operatorname{argmax}\left[\log p(X \mid \theta)\right]$$

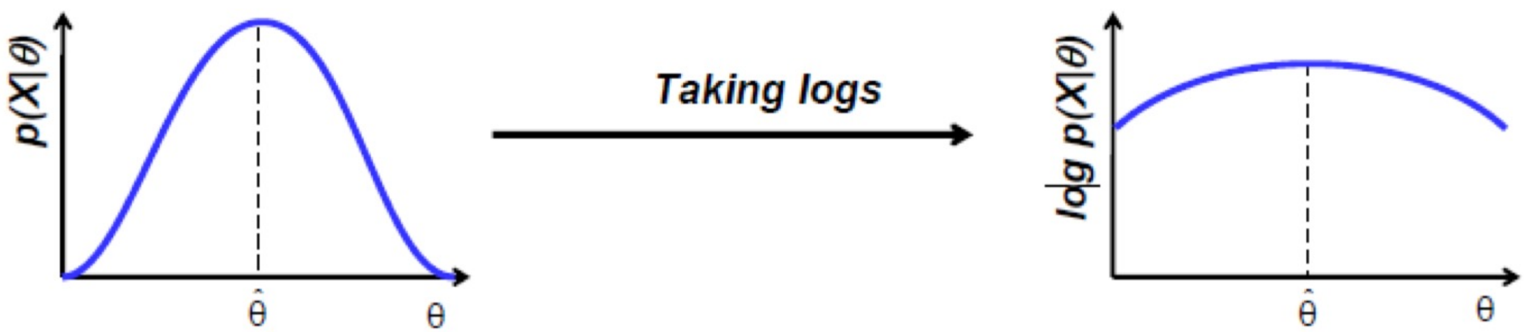


Image credit: Gutierrez-Osuna

$$\frac{1}{\sigma^2} \cdot \underbrace{\sum_{n=1}^{N} x_n - \sum_{n=1}^{N} \mu}_{\phi} = 0$$

$$\sum_{x=1} x_n = N \cdot \mu$$

Assume σ (stdev) is known:

$$\mu_{ML} = \underset{\mu}{\operatorname{argmax}} \prod_{n=1}^{N} N(x_n \mid \mu, \sigma^2)$$

μ_{ML} which maximizes the top line will also maximize log …

$$\Rightarrow \mu = \frac{1}{N} \sum_{x=1}^{n} x_n$$

N: # of data pts

$$\mu_{ML} = \operatorname{argmax} \log \prod_{n=1}^{N} N(x_n \mid \mu, \sigma^2)$$

$$= \sum_{n=1}^{N} \log N(x_n \mid \mu, \sigma^2)$$

$$= \sum_{n=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x_n-\mu)^2}{\sigma^2}\right)\right) = \underbrace{\sum \frac{1}{\sqrt{2\pi\sigma}} + \sum -\frac{(x_n-\mu)^2}{\sigma^2}}$$

$$= \sum_{n=1}^{N} \left\{ \log\left(\frac{1}{\sqrt{2\pi\sigma}}\right) - \frac{1}{2\sigma^2}(x_n-\mu)^2 \right\} = \sum \log\left(\frac{1}{\sqrt{2\pi\sigma}}\right) - \sum \frac{1}{2\sigma^2}(x_n-\mu)^2$$

$$\frac{\partial (\ldots)}{\partial \mu} = \phi + \sum \frac{1}{2\sigma^2} \cdot 2 (x_n-\mu) \cdot (-1)$$

$$\frac{1}{2\sigma^2} \cdot \left(\sum x_n - \sum \mu\right) = \phi$$

Taking the derivative with respect to μ and setting to 0,
we see that the first term does not contribute and from the second term
we find the μ as the sample mean.

$$\frac{\partial log P(\mathbf{x}|\mu, s^2)}{\partial \mu} = 0 \quad \Rightarrow \quad \boxed{\mu = \frac{1}{N}\sum_{n=1}^{N} x_n}$$

$$= \sum_{n=1}^{N}\left\{\log\left(\frac{1}{\sqrt{2\pi\sigma}}\right) - \frac{1}{2\sigma^2}(x_n - \mu)^2\right\}$$

- shown on prev. slide
- try to replicate yourself.

- If we have some prior belief about the possible values of the parameters, then we can use Maximum a Posteriori (MAP) estimate.

  - Maximum Likelihood Estimate (MLE): choose θ that maximizes probability of observed data $\mathcal{D}$

$$\widehat{\theta} = \arg\max_{\theta} P(\mathcal{D} \mid \theta)$$

  - Maximum a Posteriori (MAP) estimate: choose θ that is most probable given prior probability and the data

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} P(\theta \mid \mathcal{D}) \\ &= \arg\max_{\theta} = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}\end{aligned}$$

- MAP estimate will not be covered in more details in this course; but know the difference between the two estimates.