FYS-2021 Machine Learning

Bayes Rule and Classification (Part 1)

Slides by Stine Hansen
Guest Lecture by Elisabeth Wetzer

UiT The Arctic University of Norway

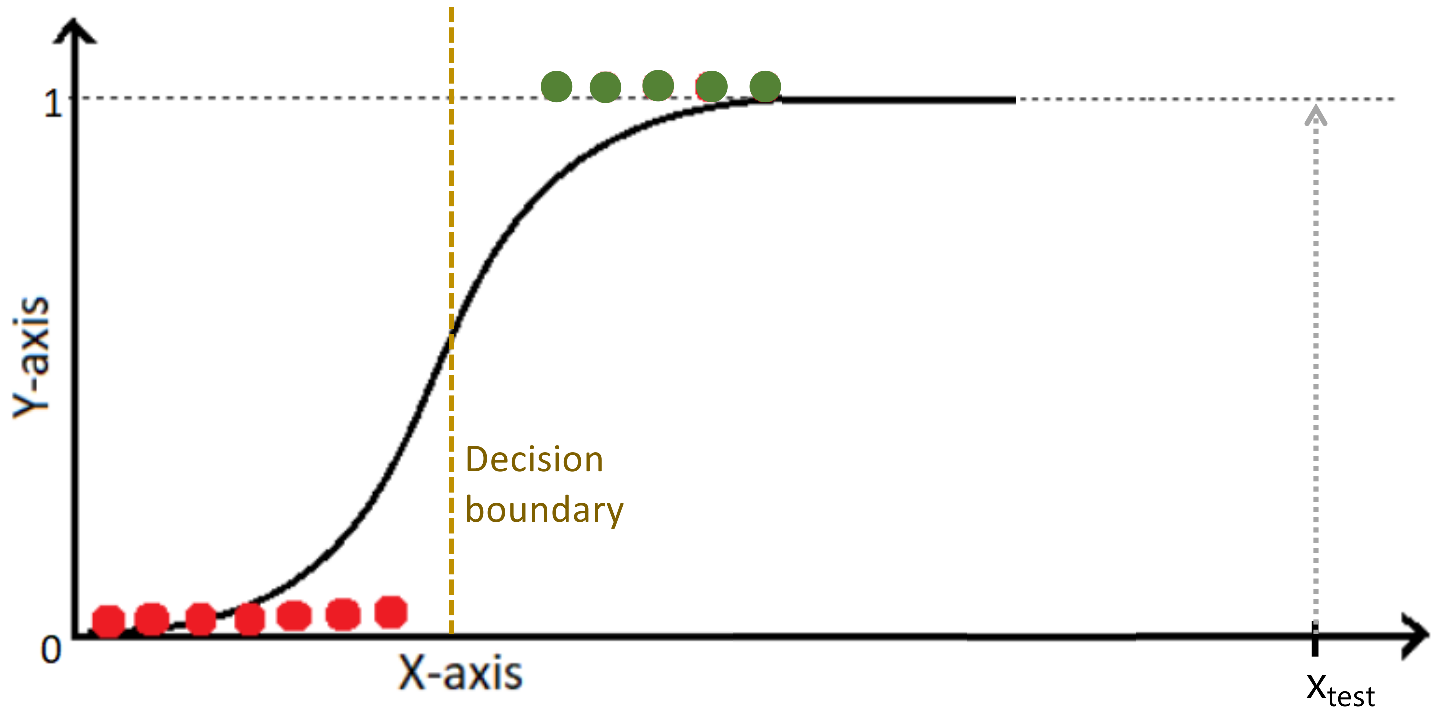
Roadmap

1. Probabilistic thinking and random variables
2. Intuition of the Bayes classifier
3. Bayes rule
4. Bayes decision rule

   Break

5. Training a Bayes Classifier
6. Maximum likelihood
7. The univariate Gaussian Bayes classifier
8. Practical connection

Why probabilistic machine learning?



$\hat{y}_{test} \approx 1$

Desirable to return a probability!

Modelling uncertainty

The real world is uncertain.

Probability is used to quantify uncertainty.

In the Bayes classifier:

- Features and class labels are treated as random variables
- Dependencies between random variables are encoded in probability distributions

Probabilistic thinking

A random variable is a function that assigns a value to each outcome of a random experiment.

A probability distribution describes how the probabilities are distributed over the possible values of a random variable.

Probabilistic thinking: Discrete random variables

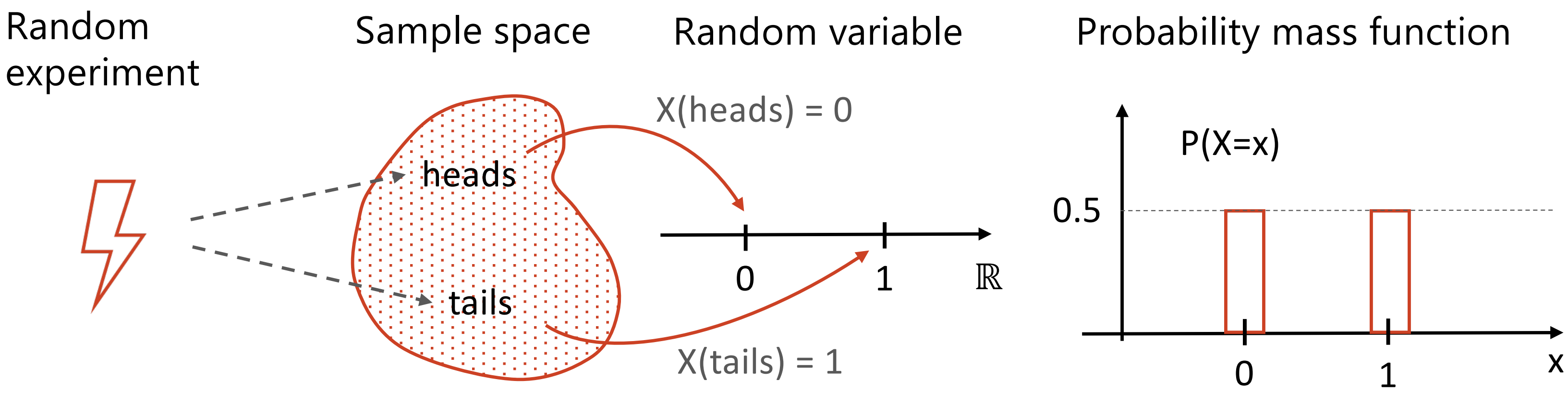
$X=\begin{cases}0 \text{ if heads} \\ 1 \text{ if tails}\end{cases}$

A random variable is a function that assigns a value to each outcome of a random experiment.

coin toss

pmf

A probability distribution describes how the probabilities are distributed over the possible values of a random variable.



Probabilistic thinking: Continuous random variables

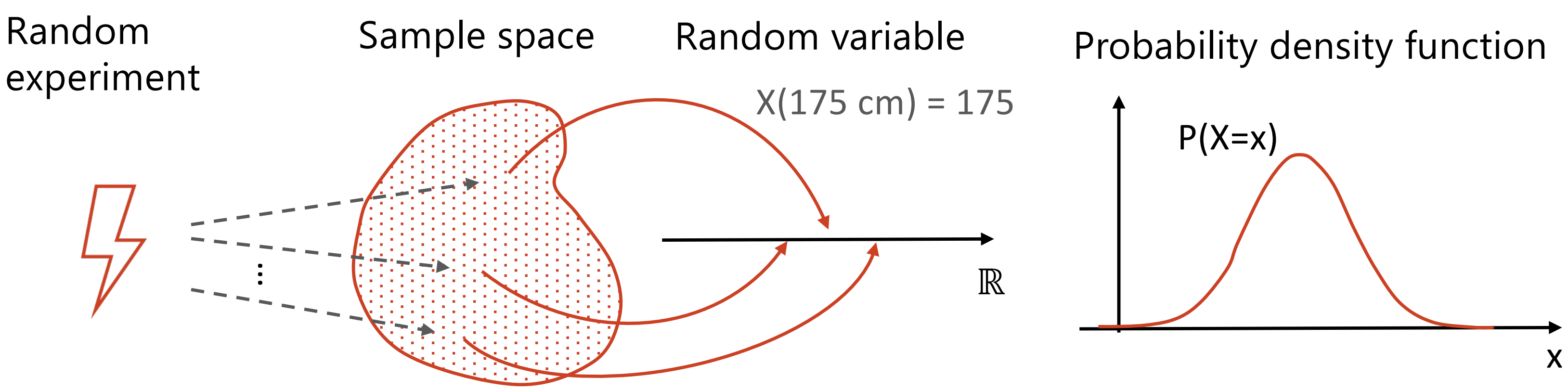
$X \in \mathbb{R}^+$

A random variable is a function that assigns a value to each outcome of a random experiment.

Measure the height of individuals

pdf

A probability distribution describes how the probabilities are distributed over the possible values of a random variable.



Probability theory

Sum rule is defined as

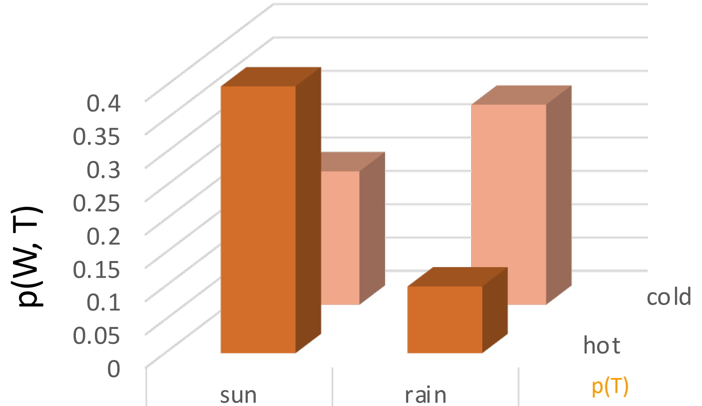
$p(X) = \sum_Y p(X,Y)$ joint distribution

Product rule is defined as

marginal distribution

$p(X,Y) = p(Y|X)p(X) = p(X|Y)p(Y)$

conditional distribution



	sun	rain	p(T)
hot	0.4	0.1	0.5
cold	0.2	0.3	0.5
p(W)	0.6	0.4	

Marginal distributions are obtained through the sum rule

$p(W) = \sum_T p(W,T)$ and $p(T) = \sum_W p(W,T)$

Conditional probabilities are computed as

$$p(W|T) = \frac{p(W,T)}{p(T)}$$

W	P(W|T=hot)
sun	0.4/0.5 = 0.8
rain	0.1/0.5 = 0.2

W	P(W|T=cold)
sun	0.2/0.5 = 0.4
rain	0.3/0.5 = 0.6

Notation

Description	Notation
Random variable	X
Realization of random variable	x
Set of N realizations	$\{x^1, x^2, \ldots, x^N\}$
Class i	C_i
Model parameters	θ

The intuition behind the Bayes classifier

height Basketball player or not

Example: Classify a realization of a random variable $X = x$ into one of two classes C_1 and C_2.

Bayes Classifier: "Assign the sample to the **most probable class, given the data**"

Decision rule:

$$x \rightarrow C_1 \quad if \quad p(C_1|X = x) > p(C_2|X = x)$$

Need to compute these conditional probabilities!

How to compute $p(C_i|X = x)$?

Use Bayes' Rule!

$$\underbrace{p(C_i|x)}_{\text{Posterior}} = \frac{\overbrace{p(x|C_i)}^{\text{Likelihood}}\ \overbrace{p(C_i)}^{\text{Prior}}}{\underbrace{p(x)}_{\text{Evidence}}}$$

Normalizing constant!

$$p(x) = \underbrace{\sum_{C_i} p(x, C_i)}_{\text{sum rule}} = \underbrace{\sum_{C_i} p(x|C_i)p(C_i)}_{\text{product rule}}$$

$$p(x|C_i)p(C_i) \underbrace{=}_{\text{product rule}} p(x, C_i) \underbrace{=}_{\text{product rule}} p(C_i|x)p(x)$$

How to compute $p(C_i|X = x)$?

Example: Compute the probability of having a disease given a positive test.

$C = \{\text{disease}, \text{not disease}\}$
$X \in \{1, 0\}$ (positive/negative test) discrete random variable

Want to compute

$$\underset{\text{Posterior}}{p(C_i|x)} = \frac{\overset{\text{Likelihood}}{p(x|C_i)}\ \overset{\text{Prior}}{p(C_i)}}{\underset{\text{Evidence}}{p(x)}}$$

$$p(x) = \underbrace{\sum_{C_i} p(x, C_i)}_{\text{sum rule}} = \underbrace{\sum_{C_i} p(x|C_i)p(C_i)}_{\text{product rule}}$$

Posterior $\mathrm{p}(\text{disease}|x = 1) = \frac{p(x=1|\text{disease})p(\text{disease})}{p(x=1)} = \frac{0.8 * 0.004}{0.004 * 0.8 + 0.996 * 0.1} = 0.031$ ← If the test is positive, the chance of disease is 3 %

Likelihood $p(x = 1|\text{disease}) = 0.8$ (sensitivity of test is known to be 80 %)

Prior $p(\text{disease}) = 0.004$ (0.4 % of the population gets the disease)

Evidence $p(x = 1) = p(\text{disease})p(x = 1|\text{disease}) + p(\text{not disease})p(x = 1|\text{not disease})$
$= 0.004 * 0.8 + 0.996 * 0.1$
(FP rate of test is known to be 10 %)

Example from: Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press.

Bayes Classifier: Decision rule

$$p(C_i|x) = \frac{\overset{\text{Likelihood}}{p(x|C_i)}\ \overset{\text{Prior}}{p(C_i)}}{\underset{\text{Evidence}}{p(x)}}$$

$$x \to C_1 \quad if \qquad p(C_1|x) > p(C_2|x)$$

$$\frac{p(x|C_1)\,p(C_1)}{p(x)} > \frac{p(x|C_2)\,p(C_2)}{p(x)}$$

$$p(x|C_1)\,p(C_1) > p(x|C_2)\,p(C_2)$$

$$\boxed{x \to C_i \quad if \quad p(x|C_i)\,p(C_i) > p(x|C_j)\,p(C_j)\ \forall\, j \neq i}$$

Bayes classifier

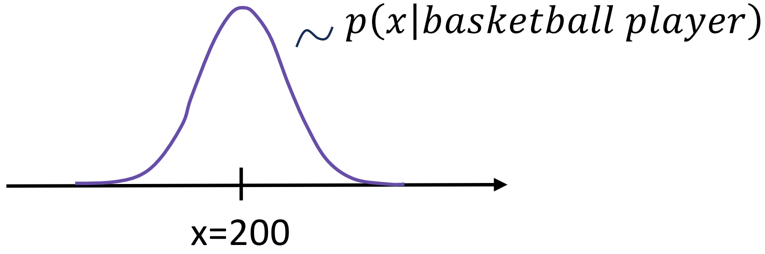
Example: Classify whether a person is a basketball player based on height.

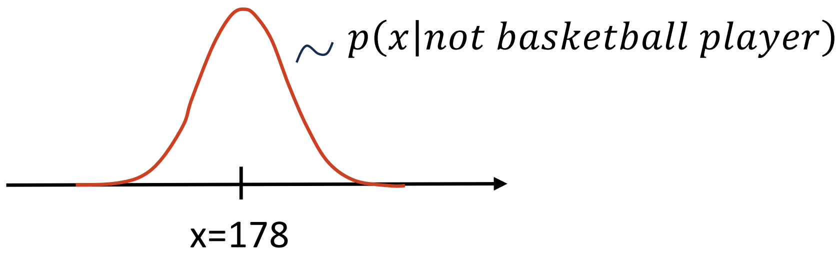
C = {basketball player, not basketball player}
$X \in \mathbb{R}$ (height in cm) continuous random variable

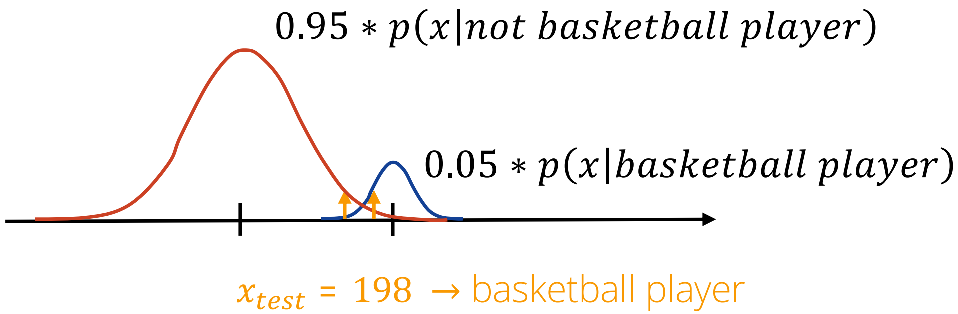
- Estimate likelihoods, $p(x|C_i)$
- Estimate priors, $p(C_i)$

$$p(basketball\ player) = 0.05$$

$$p(not\ basketball\ player) = 0.95$$





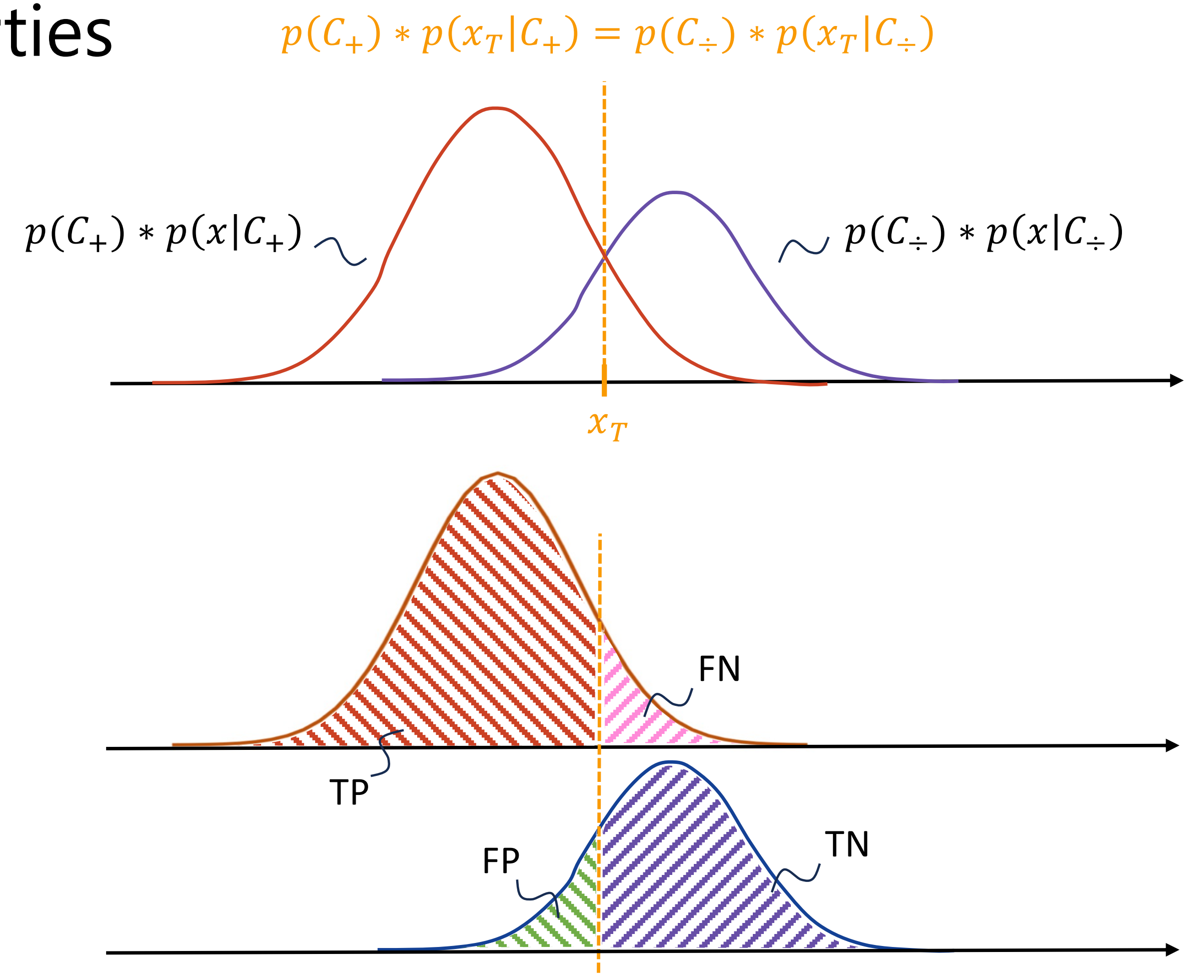


Bayes classifier: Properties

- Threshold where $p(C_+|x) = p(C_\div|x)$

- $p(error) = \int_{x_T}^{\infty} p(C_+) * p(x|C_+)dx + \int_{-\infty}^{x_T} p(C_\div) * p(x|C_\div)dx$

- Probability of errors is minimized if we have the true $p(x|C_i)$ and $p(C_i)$!



Roadmap

1. Probabilistic thinking and random variables
2. Intuition of the Bayes classifier
3. Bayes rule
4. Bayes decision rule

   Break

5. Training a Bayes Classifier
6. Maximum likelihood
7. The univariate Gaussian Bayes classifier
8. Practical connection

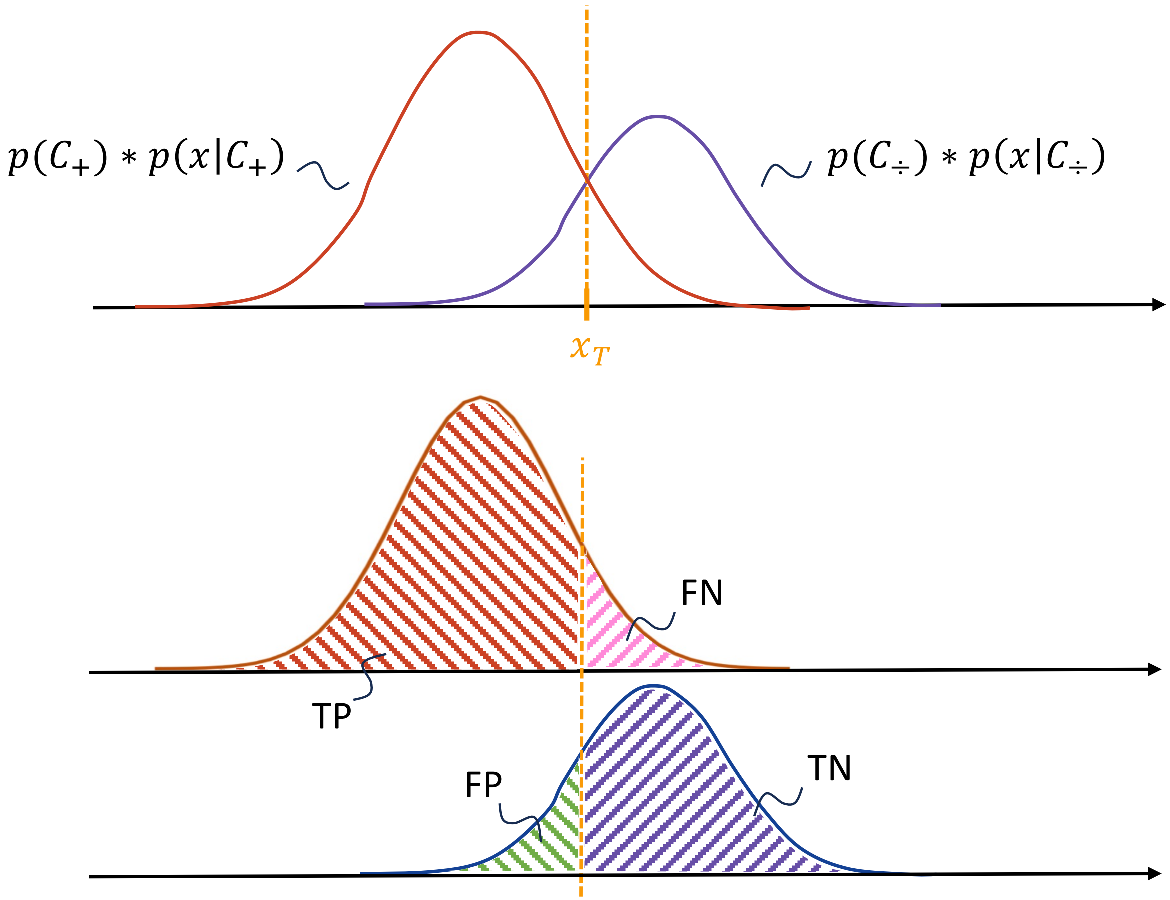
Bayes classifier: Properties

- Threshold where $p(C_+|x) = p(C_\div|x)$

- $p(error) = \int_{x_T}^{\infty} p(C_+) * p(x|C_+)dx + \int_{-\infty}^{x_T} p(C_\div) * p(x|C_\div)dx$

- Probability of errors is minimized if we have the true $p(x|C_i)$ and $p(C_i)$!

$p(C_+) * p(x_T|C_+) = p(C_\div) * p(x_T|C_\div)$



Need to estimate $p(C_i)$ and $p(x|C_i)$ from data

To train a Bayes classification model = To estimate $p(C_i)$ and $p(x|C_i)$ from training data $\mathcal{D} = \{(x^i, y^i)\}_{i=1}^N$!

Scaling Shape/position

Prior class probabilities, $p(C_i)$

are estimated as the frequency of class C_i in the training data:

$$p(C_i) = \frac{\#\mathcal{D}\{y^i = C_i\}}{N}$$

Number of training samples belonging to class C_i

Likelihood terms, $p(x|C_i)$

are estimated via **maximum likelihood**!

Maximum likelihood estimation (MLE)

Estimating the parameter values θ of a statistical model that maximize the likelihood of the observed data

In practice:

1. Assume some parameterized distribution $p_\theta(x|C_i)$
   can be Gaussian, Laplacian, Bernoulli, etc. depending on the type of features

2. Estimate the distribution's parameter(s) as

$$\hat{\theta}_{MLE} = arg\max_\theta \ell(\theta|\mathcal{D}) \equiv arg\max_\theta p_\theta(x^1, \dots, x^N|C_i)$$

iid

$$= arg\max_\theta \prod_{j=1}^{N} p_\theta(x^j|\, C_i)$$

Ind. variables: P(A, B) = P(A)*P(B)

$$= arg\max_\theta \underbrace{\sum_{j=1}^{N} \log p_\theta(x^j|C_i)}_{\text{log likelihood, } L(\theta|\mathcal{D})}$$

log(a*b) = log(a) + log(b)

Maximum likelihood estimation (MLE)

log likelihood, $L(\theta|\mathcal{D})$

$$\hat{\theta}_{MLE} = \underset{\theta}{argmax} \sum_{j=1}^{N} \log p_\theta(x^j|C_i) \quad \rightarrow \text{solve:} \qquad \frac{\partial}{\partial\theta} L(\theta|\mathcal{D}) = 0$$

Example: Univariate gaussian distribution, $p_\theta(x|C_i) \sim \mathcal{N}(\mu, \sigma)$ Here: θ=[μ, σ]

Assuming we have N_i observed samples from class C_i, $\{x^1, x^2, \dots, x^{N_i}\}$

$$L(\theta|\mathcal{D}) = \sum_{j=1}^{N_i} \log p_\theta(x^j|C_i)$$

$$= \sum_{j=1}^{N_i} \log\left[\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x^j-\mu}{\sigma}\right)^2}\right]$$

$$= -N_i \log\left(\sqrt{2\pi}\right) - N_i \log(\sigma) - \frac{1}{2}\sum_{j=1}^{N_i}\left(\frac{x^j-\mu}{\sigma}\right)^2$$

$$\frac{\partial}{\partial\mu} L(\theta|\mathcal{D}) = 0 \quad \rightarrow \quad \hat{\mu}_{MLE} = \frac{1}{N_i}\sum_{j=1}^{N_i} x^j$$

$$\frac{\partial}{\partial\sigma} L(\theta|\mathcal{D}) = 0 \quad \rightarrow \quad \hat{\sigma}^2_{MLE} = \frac{1}{N_i}\sum_{j=1}^{N_i}\left(x^j - \mu\right)^2$$

(Univariate) Gaussian Bayes classifier

Putting it all together

Assuming we have N training samples from class C_+, $\{x^1, x^2, \ldots, x^N\}$, and M training samples from class $C_\div$, $\{x^1, x^2, \ldots, x^M\}$, and that $p(x|C_i) \sim \mathcal{N}(\mu_i, \sigma_i)$, $\; i = \{+, \div\}$.

Training:

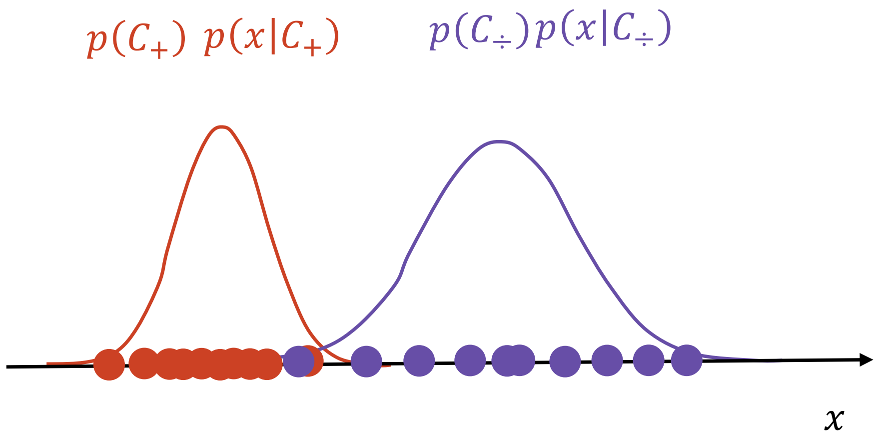
MLEs for likelihood terms are given by

$$\hat{\mu}_+ = \frac{1}{N}\sum_{j=1}^{N} x^j \qquad \hat{\mu}_\div = \frac{1}{M}\sum_{j=1}^{M} x^j$$

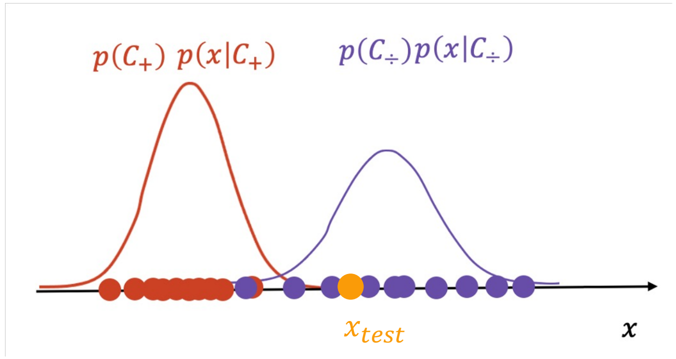
$$\hat{\sigma}_+^2 = \frac{1}{N}\sum_{j=1}^{N}\left(x^j - \mu\right)^2 \qquad \hat{\sigma}_\div^2 = \frac{1}{M}\sum_{j=1}^{M}\left(x^j - \mu\right)^2$$

Prior class probability estimates are given by

$$\hat{p}(C_+) = \frac{N}{N+M} \qquad \hat{p}(C_\div) = \frac{M}{N+M}$$



(Univariate) Gaussian Bayes classifier



Test time (inference)

$$\boxed{\begin{array}{c} x \ \rightarrow C_i \quad if \\ p(x|C_i)\, p(C_i) > p(x|C_j)\, p(C_j)\ \forall\, j \neq i \end{array}}$$

Decision rule

When we have some test sample x_{test} that we want to classify, we need to evaluate

$$g_i(x_{test}) = \ p(x_{test}|C_i)\, p(C_i)$$

(discriminant function)

for all i. Then chose class C_i if

$$g_i(x_{test}) = \ \max_k g_k(x_{test})$$

$$g_+(x_{test}) = \left(\frac{1}{\sqrt{2\pi}\,\hat{\sigma}_+} e^{-\frac{1}{2}\left(\frac{x_{test}-\hat{\mu}_+}{\hat{\sigma}_+}\right)^2}\right)\left(\frac{N}{N+M}\right)$$

$$g_-(x_{test}) = \left(\frac{1}{\sqrt{2\pi}\,\hat{\sigma}_\div} e^{-\frac{1}{2}\left(\frac{x_{test}-\hat{\mu}_\div}{\hat{\sigma}_\div}\right)^2}\right)\left(\frac{M}{N+M}\right)$$

Quiz: Training a Bayes classifier

Assume a three-class classification problem where the likelihoods are Gaussians. How many parameters do you have to estimate to train the univariate Bayes classifier?

A: 3 parameters

B: 6 parameters

C: 9 parameters

D: It depends on the dataset

Answer: 3 classes * (prior + mean + std) = 9 parameters

Practical connection

0. [If not done] Split data into train/test

On training data:

1. Sort samples according to labels (class)
2. For each class: Compute estimates for class prior as class frequency
3. For each class: Assume distribution of likelihood and compute corresponding $\hat{\theta}_{MLE}$

On test data:

1. For each class: Compute discriminant function
2. Assign samples to class with maximum discriminant function
3. Compare predictions to labels and compute your confusion matrix

Roadmap

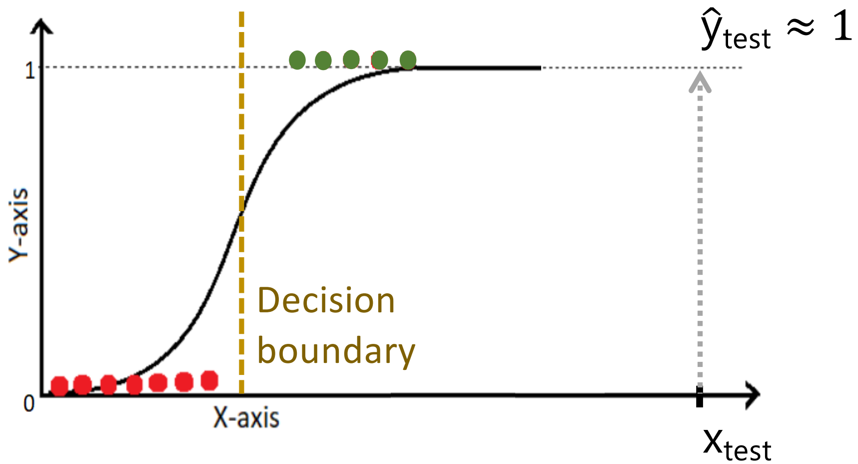
1. Probabilistic thinking and random variables
2. Intuition of the Bayes classifier
3. Bayes rule
4. Bayes decision rule

    Break

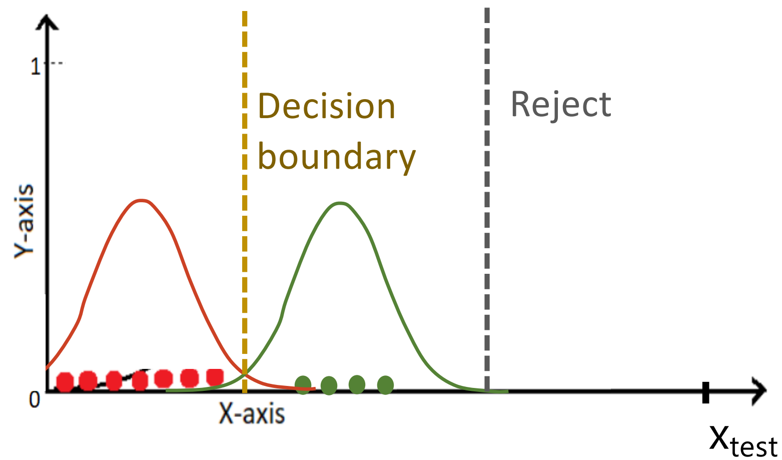
5. Training a Bayes Classifier
6. Maximum likelihood
7. The univariate Gaussian Bayes classifier
8. Practical connection

Reject option

Logistic regression



Bayes classifier w/ reject option



$$x \rightarrow C_i \quad if \quad p(C_i|x) > p(C_j|x) \forall\, j \neq i$$

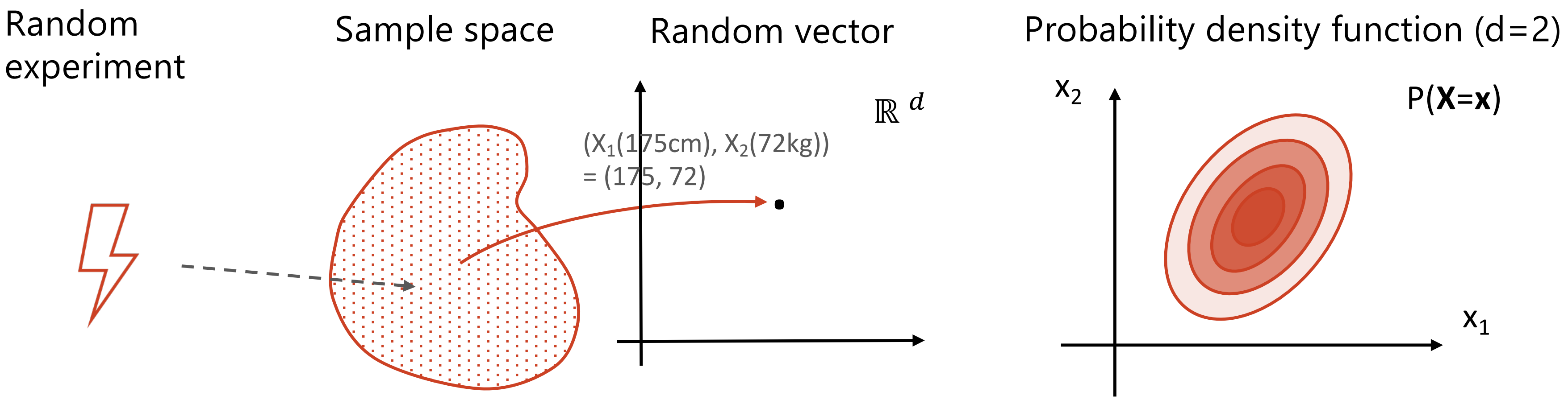
$$\text{and } p(C_i|x) > 1 - \lambda$$

reject otherwise

Conclusion

- Supervised classification model
- Bayes classifier for univariate data
  - Assume distribution for the likelihoods
  - Learn class priors and MLE parameters for each class
- Closed form solutions to MLE (no iterative optimization)
- Possible to reject samples during inference

Thursday: From univariate to multivariate data



Video by 3Blue1Brown

- https://www.youtube.com/watch?v=HZGCoVF3YvM