Istituto Dalle Molle di studi sull'intelligenza artificiale
USI/SUPSI
IDSIA

Bayes' Theorem and Estimation Theory

Machine Learning 2019

Michael Wand, Jürgen Schmidhuber, Cesare Alippi

TAs: Robert Csordas, Krsto Prorokovic, Xingdong Zou, Francesco Faccio, Louis Kirsch

based on slides by Jan Unkelbach (only Bayes part)

Introduction

- We have revisited basic probability theory
- In particular, we have gotten to know *conditional probabilities*
- They form the basis of *Bayes' theorem* – fundamental in
  - probability theory and statistics
  - logical reasoning
  - machine learning
  - …
- Today we look at Bayes theorem
- … relate it to the theory/practice of estimation
- … and derive a well-known classifier (the Gaussian classifier)

Bayes' Theorem

Recap: Conditional Probability

Assume *events* A and B. We had defined

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

or, using random variable formulation with r.v. X and Y (⋏ means “and”):

$$P(Y \in M | X \in N) = \frac{P(Y \in M \curlywedge X \in N)}{P(X \in N)}$$

Recap: Frequentist Statistics

	Definition	Frequentist Interpretation
$P(A)$	Probability that event A happens	Assume we repeat an experiment many times. $P(A)$ is the fraction of trials in which A has happened.
$P(B \cap A)$	Probability that both A and B happen	Assume we repeat an experiment many times. $P(B \cap A)$ is the fraction of trials in which A and B have happened.
$P(B \mid A)$	Probability that B happens, given that A has “already” happened / is known to have happened	Assume we repeat an experiment many times. $P(B\|A)$ is the fraction of trials in which happened B and A, *out of those* where A has happened

Bayes' Theorem

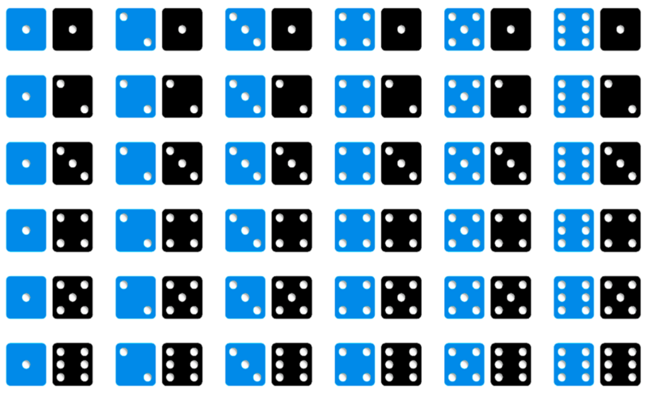
- Bayes' Theorem relates conditional probabilities in different "directions"!

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

Thomas Bayes
(ca. 1701 – 1761),
philosopher, statistician,
Presbyterian minister
(source: Wikipedia)

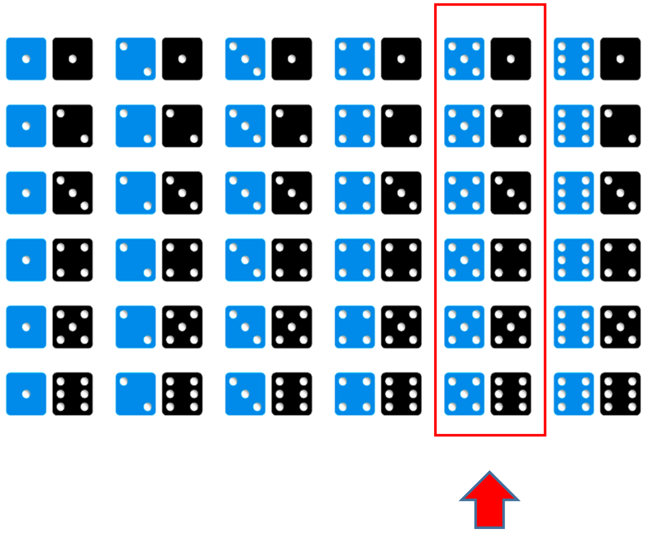
Bayes' Theorem: Simple Example

- Assume that we roll two (fair) dice, as last week.
  - Event A: First die shows a 5 (P(A)=1/6)
  - Event B: Second die shows a 3 (P(B)=1/6)
  - Event C: The sum of both dice is 10 (P(C)=1/12)
- $P(A \cap B) = 1/36$, $P(A \cap C) = 1/36$, $P(B \cap C)=0$
- Also remember that only A and B are independent.



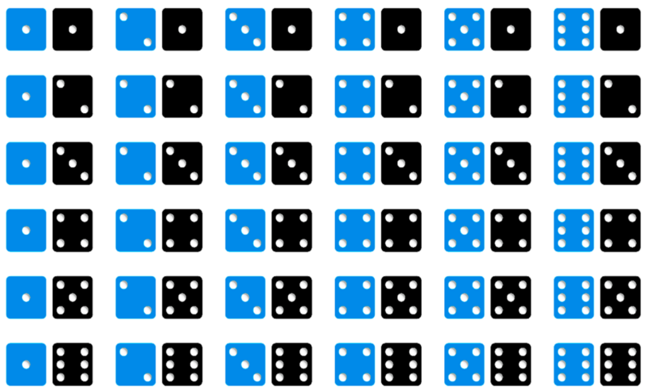
Bayes' Theorem: Simple Example

- Assume that we roll two (fair) dice, as last week.
  - Event A: First die shows a 5 ($P(A)=1/6$)
  - Event B: Second die shows a 3 ($P(B)=1/6$)
  - Event C: The sum of both dice is 10 ($P(C)=1/12$)
- $P(A \cap B) = 1/36$, $P(A \cap C) = 1/36$, $P(B \cap C)=0$.
- Also remember that only A and B are independent.
- Last time we computed $P(C|A) = 1/6$.



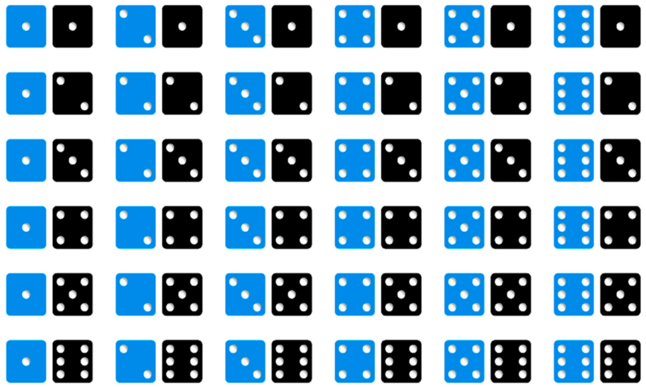
Bayes' Theorem: Simple Example

- Assume that we roll two (fair) dice, as last week.
  - Event A: First die shows a 5 (P(A)=1/6)
  - Event B: Second die shows a 3 (P(B)=1/6)
  - Event C: The sum of both dice is 10 (P(C)=1/12)
- $P(A \cap B) = 1/36$, $P(A \cap C) = 1/36$, $P(B \cap C)=0$.



- $P(C|A) = 1/6$. But what is $P(A|C)$?

Bayes’ Theorem: Simple Example



- Assume that we roll two (fair) dice, as last week.
  - Event A: First die shows a 5 (P(A)=1/6)
  - Event B: Second die shows a 3 (P(B)=1/6)
  - Event C: The sum of both dice is 10 (P(C)=1/12)
- P(A∩B) = 1/36, P(A∩C) = 1/36, P(B∩C)=0.

- P(C|A) = 1/6. But what is P(A|C)?
- Bayes’ Theorem:

$$P(A|C) = \frac{P(C|A) \cdot P(A)}{P(C)} = \frac{1/6 \cdot 1/6}{1/12} = \frac{12}{36} = \frac{1}{3}$$

Bayes' Theorem: Surprising Example

(from medicine, and quite realistic)

- C: a patient has breast cancer
- X: positive mammogram is observed in screening
- P(C) = 0.01: *prior probability* that a randomly chosen woman has breast cancer *without* knowing the results of the exam
- P(¬C) = 0.99: prior probability that a randomly chosen woman is healthy
- P(X|C) = 0.8: probability of diagnosing existing cancer (*sensitivity*)
- P(X|¬C) = 0.1: probability of a *false positive*

Question: what is P(C|X), i.e. the probability of actually having cancer given the mammogram is positive

Bayes' Theorem: Surprising Example

Apply Bayes' Theorem:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

Bayes' Theorem: Surprising Example

Apply Bayes' Theorem:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

We can compute

$P(X) = P(X \cap C) + P(X \cap \neg C) = P(X|C)P(C) + P(X|\neg C)P(\neg C)$
$= 0.8 \cdot 0.01 + 0.1 \cdot 0.99 = 0.107$

by partitioning the event space!

Bayes' Theorem: Surprising Example

Apply Bayes' Theorem:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

P(X) = 0.107. So,

$$P(C|X) = \frac{0.8 \cdot 0.01}{0.107} \approx 0.075 = 7.5\%$$

a *small* probability of actually having cancer!

Intuitive explanantion: we are detecting a **rare event** with a non-perfect test. The probability of a false positive result is larger than picking a cancer patient.

What do we learn?

- Incorporating prior knowledge is important!
- If an event which plays a role in our argumentation is rare, we may get unintuitive results!
- Remember that we had $P(X|C) \approx 1$, which is what we intuitively hope of a diagnostic test
  - but $P(C|X)$ was actually very small since $P(C)$ and $P(X)$ had different orders of magnitude, and

$$P(C|X) = P(X|C) \cdot \frac{P(C)}{P(X)}$$

Final definitions

Likelihood of A given B

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

Prior probability of B

Posterior probability
of B given A

Evidence (probability of A)

Final definitions (RV form)

Istituto Dalle Molle di studi sull'intelligenza artificiale
USI/SUPSI IDSIA

Likelihood of A given B

Prior probability of B

$$\boxed{P(B=y|A=x) = \frac{P(A=x|B=y)\cdot P(B=y)}{P(A=x)}}$$

Posterior probability of B given A

Evidence (probability of A)

Elements of Estimation Theory

Estimation Theory

- Definition: Estimate the values of parameters based on measured empirical data (which has a probabilistic component)
  - Probabilistic component often emerges by sampling
  - Example: We wish to estimate the number of smokers in Lugano
    - walk around and ask people (let's say, around 2000)
    - (how to make the sample representative??)
    - now we get an estimate which hopefully *roughly* reflects the true fraction of smokers
    - but obviously we get a slightly different result if we ask 2000 different people

Estimation Theory

- Definition: Estimate the values of parameters based on measured empirical data (which has a probabilistic component)
  - Probabilistic component often emerges by sampling
  - Example: We wish to estimate the number of smokers in Lugano
    - walk around and ask people (let’s say, around 2000)
    - (how to make the sample representative??)
    - now we get an estimate which hopefully *roughly* reflects the true fraction of smokers
    - but obviously we get a slightly different result if we ask 2000 different people

  - Probabilistic component can also emerge by incomplete information
    - do we know that everybody answers truthfully?

Estimation Theory

- We see that an estimate of any value is *probabilistic!*
  - We can compute its probabilistic properties, e.g. expected error, bias…
  - Prior knowledge plays an integral role!

Estimation Theory

- We see that an estimate of any value is *probabilistic!*
  - We can compute its probabilistic properties, e.g. expected error, bias…
  - Prior knowledge plays an integral role!

Connection to Bayes!

Maximum Likelihood Estimator

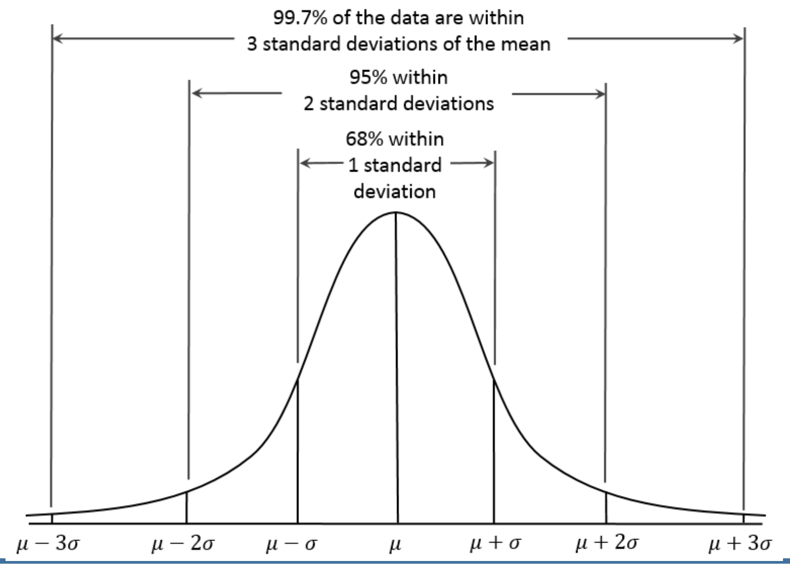
- Let us consider a simple example with animals (from Wikipedia):
  - we are interested in the heights of adult female penguins
  - no way to measure the entire population
  - but we can measure, let's say, the height of 200 penguins

Maximum Likelihood Estimator

- Let us consider a simple example with animals (from Wikipedia):
  - we are interested in the heights of adult female penguins
  - no way to measure the entire population
  - but we can measure, let's say, the height of 200 penguins

- Now comes the assumption we have to make!
  - Let us assume that the heights are normally distributed with mean μ and variance σ.

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



Maximum Likelihood Estimator

- Let us say we have measured a sample of penguin heights $X=\{x_1, \ldots x_N\}$ (N=200).
- Now we need to estimate μ (and possibly σ), based on our sample.
- How do we do that?

Maximum Likelihood Estimator

- Let us say we have measured a sample of penguin heights $X=\{x_1, \ldots x_N\}$ ($N=200$).
- Now we need to estimate μ (and possibly σ), based on our sample.
  - Assuming that we *have* μ and σ, we can compute P(X| μ, σ).
  - Simple idea: *maximize* P(X| μ, σ) over all possible values of μ, σ.
  - This can be done analytically, and the result for μ is the sample mean:

$$\hat{\mu}_{ML} = \frac{1}{N}\sum_{n}^{N} x_n$$

  - For the variance, we get a similar result:

$$\hat{\sigma}_{ML}^2 = \frac{1}{N}\sum_{n}^{N} (x_n - \mu)^2$$

The Gaussian Classifier

- We can now estimate parameters of a sample, under the assumption of Gaussianity.
- We can use this to create a simple parametric classifier: The *Gaussian Classifier.*

The Gaussian Classifier

- For each class, *assume it follows a Gaussian distribution*
- Estimate mean and variance for each class c: $\mathcal{N}_c = \mathcal{N}(\mu_c, \sigma_c)$

This amounts to the training of the classifier. We save the computed means and variances and do not need the samples any more.

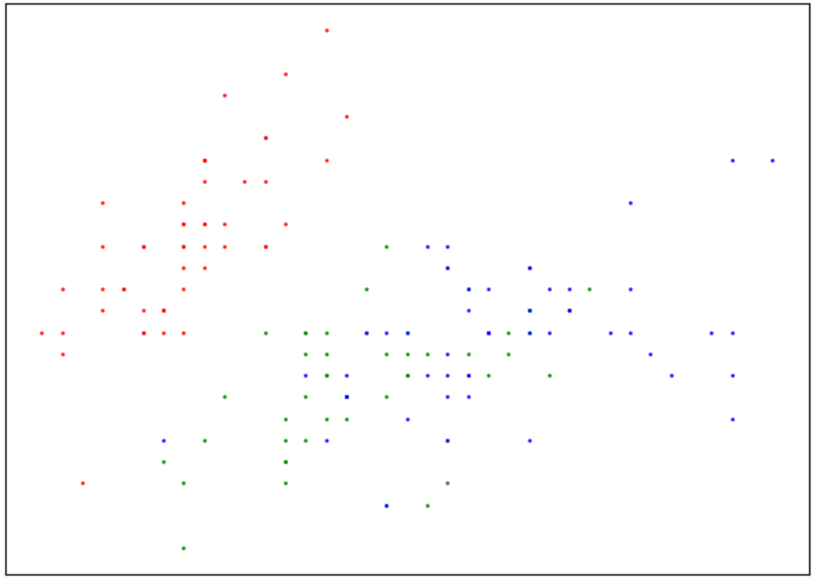
- In order to classify a new data point x, compute $\mathcal{N}_c$ for each class c
- Result: $\hat{c} = \mathrm{argmax}_c \mathcal{N}_c(x)$
- (Remark: We normally have Gaussians in *multiple* dimensions, however the principle is the same)

Gaussian Classifier Example: Iris Dataset

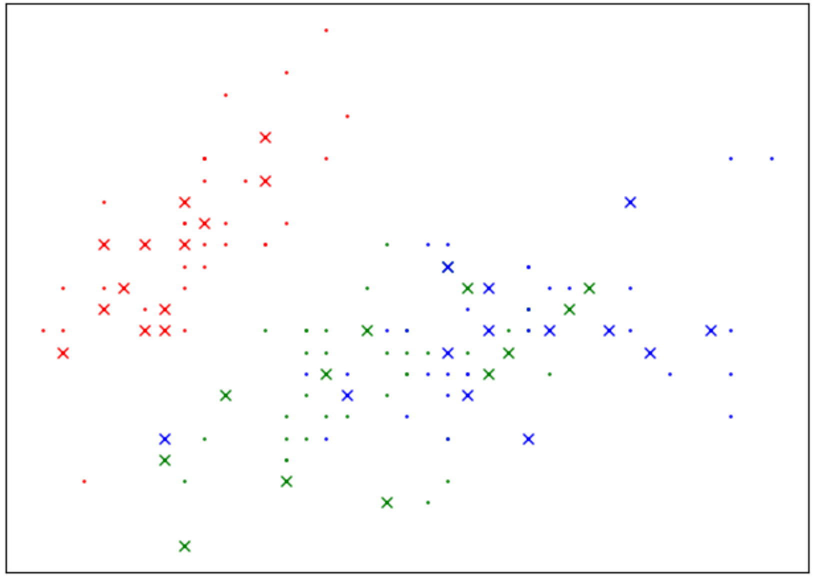
USI/SUPSI
Istituto Dalle Molle di studi sull'intelligenza artificiale
IDSIA

- The Iris dataset (Fisher, 1936)
- Three species of the Iris flower, four features related to their size, goal: distinguish the species!
- Plots show first two features

Training data

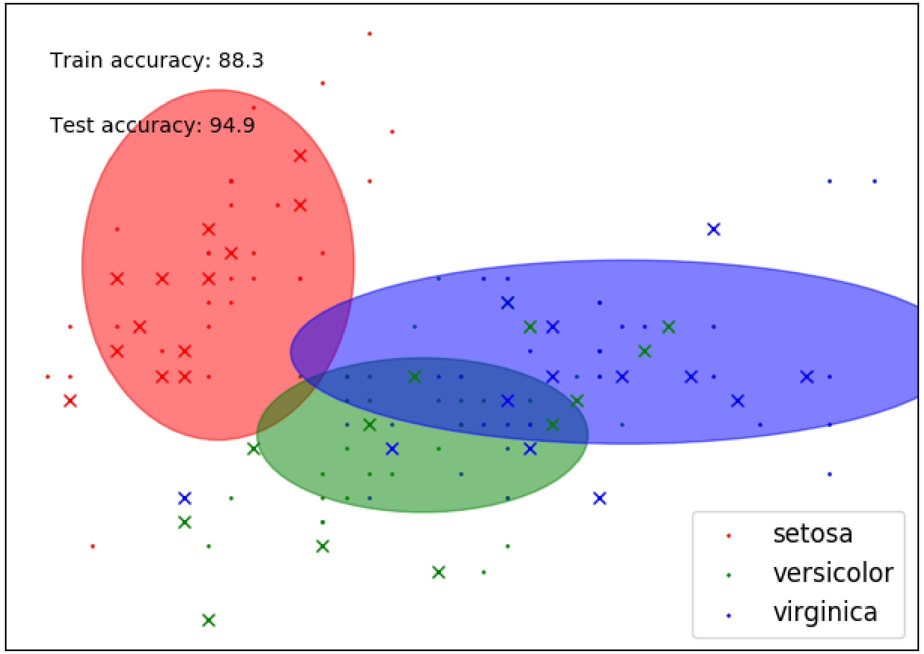


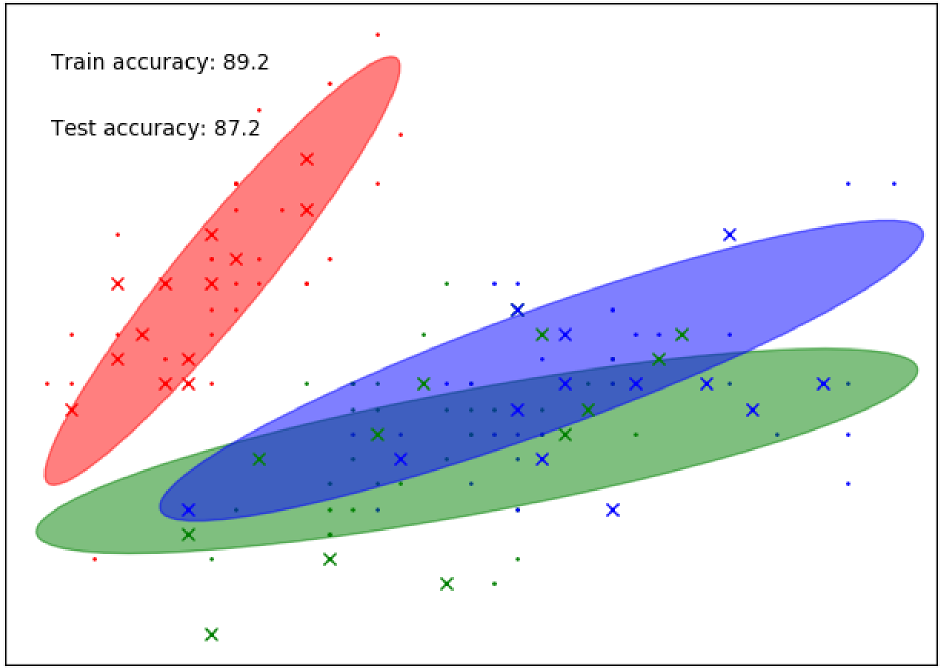
Training and test data



Gaussian Classifier Example: Iris Dataset

- Estimated Gaussians with and without considering covariance





- In practice, one Gaussian per class is not enough - requires *mixtures* of Gaussians, which we do not cover right now

Estimators

Remember that there are *many* ways to estimate a parameter from a sample!

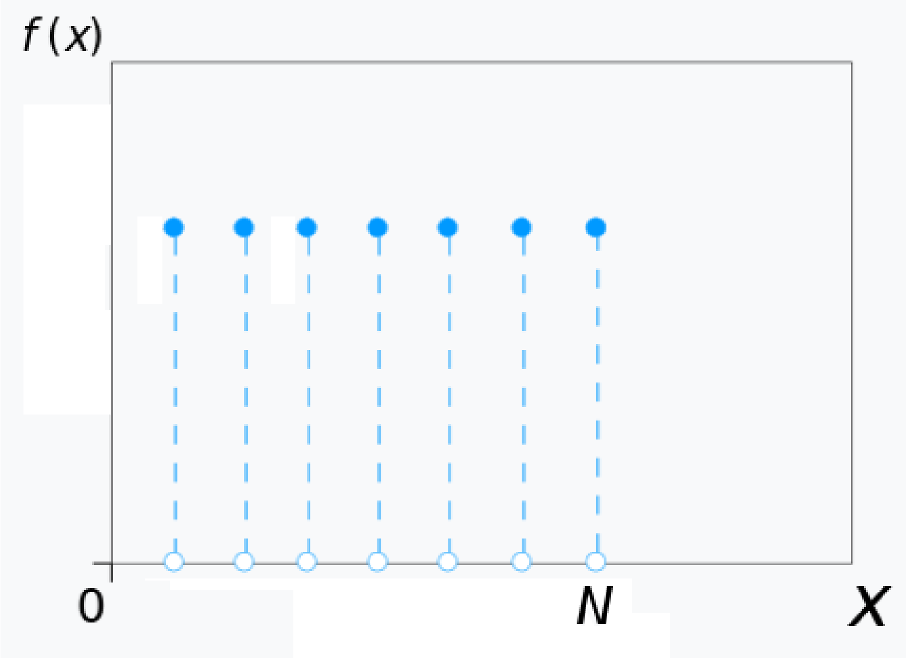
Describe estimators using statistical properties:

- *unbiasedness* – the expectation of the estimator should equal the value to be estimated
  - maximum likelihood variance estimation is *not* unbiased
- *consistency* – the more samples we use, the better our estimation gets
- *variance* – the variance of the estimate, assuming we estimate many times with different sample sets
  - low variance is typically desired

Estimation – Complex Example

Istituto Dalle Molle di studi sull'intelligenza artificiale
USI/SUPSI IDSIA

Assume a uniform distribution (from 1 to N) with unknown N. How would you estimate N based on samples x_1, …, x_L?



The maximum likelihood estimator is $\hat{N}_{max} = \max x_i =: m$.

But it can be shown that this estimator is *biased* – it underestimates the true maximum.

The unbiased solution is

$$\hat{N}_{unbiased} = m + \frac{m}{L} - 1$$

Maximum A Posteriori Estimator

- Why is maximum likelihood estimation *not* Bayesian?

➢ Because we do not take prior knowledge into account!

- Bayesian estimation always assumes that we have some prior knowledge of our parameter (let's call it θ), and that we update this prior knowledge with observations.

Likelihood of A given θ

Prior probability of θ

$$\boxed{P(\theta|A) = \frac{P(A|\theta)\cdot P(\theta)}{P(A)}}$$

Evidence (probability of A)

Posterior probability of θ given A

Maximum A Posteriori Estimator

- We do just one example: Estimation of a sample mean (so, θ={μ})
- Assume we have a sequence of independent samples x_1, …, x_N, which we assume follow a Gaussian distribution: $X \sim \mathcal{N}(\mu_s, \sigma_s)$
- For the mean, we assume a Gaussian *prior* distribution: $\mu_s \sim \mathcal{N}(\mu_0, \sigma_0)$

Maximum A Posteriori Estimator

- We do just one example: Estimation of a sample mean (so, $\theta=\{\mu\}$)
- Assume we have a sequence of independent samples $x_1, \ldots, x_N$, which we assume follow a Gaussian distribution: $X \sim \mathcal{N}(\mu_s, \sigma_s)$
- For the mean, we assume a Gaussian *prior* distribution: $\mu_s \sim \mathcal{N}(\mu_0, \sigma_0)$
- It can be shown that the *Maximum a Posteriori (MAP)* estimate for μ_s, after observing the samples, is

$$\hat{\mu}_s^{MAP} = \frac{\sigma_0^2 (\sum_i x_i) + \sigma_s^2 \mu_0}{\sigma_0^2 n + \sigma_s^2}$$

… a weighted linear interpolation between "old" mean and "new" mean

Maximum A Posteriori Estimator

- We do just one example: Estimation of a sample mean (so, θ={μ})
- Assume we have a sequence of independent samples x_1, …, x_N, which we assume follow a Gaussian distribution: $X \sim \mathcal{N}(\mu_s, \sigma_s)$
- For the mean, we assume a Gaussian *prior* distribution: $\mu_s \sim \mathcal{N}(\mu_0, \sigma_0)$
- It can be shown that the *Maximum a Posteriori (MAP)* estimate for μ_s, after observing the samples, is

$$\hat{\mu}_s^{MAP} = \frac{\sigma_0^2 (\sum_i x_i) + \sigma_s^2 \mu_0}{\sigma_0^2 n + \sigma_s^2}$$

  … a weighted linear interpolation between "old" mean and "new" mean
- In the case of $\sigma_0 \to \infty$, we get a less and less informative prior, and the MLE estimator as limit

Maximum A Posteriori Estimator

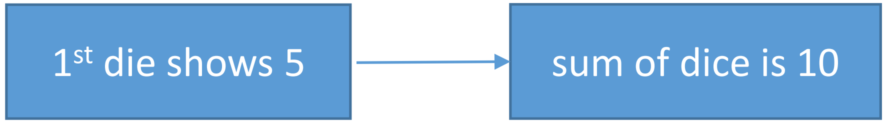
- We have seen that incorporating prior information gives very natural estimation results (ok, we have seen it from one example)
- There are different ways to obtain prior information:
  - From a related problem
  - From subjective judgement
  - From theoretical considerations
  - From prior experiments (on related data)

Final Remarks

- We distinguish *Point Estimates* from *Interval Estimates*
  - …which exist both in frequentist and Bayesian statistics
  - *Confidence/Credible Intervals* related to statistical validation
- Particularly in the case of Bayesian statistics, remember that we had a *prior* distribution for our parameter
  - … so after estimation with a random sample, we should logically get a *posterior* distribution
  - Bayesian theory and practice deals with this as well
  - … but we do not cover it here

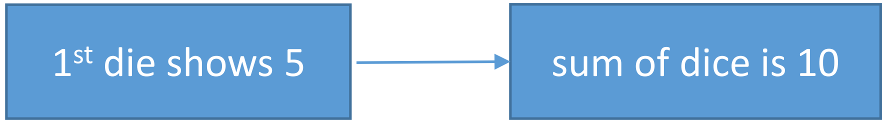
Final Remarks

- Bayes' Theorem fundamentally deals with conditional probabilities

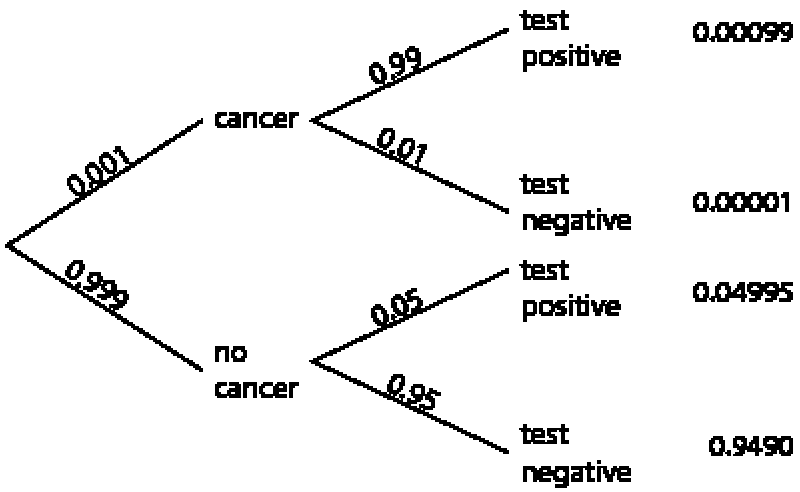


Final Remarks

- Bayes’ Theorem fundamentally deals with conditional probabilities

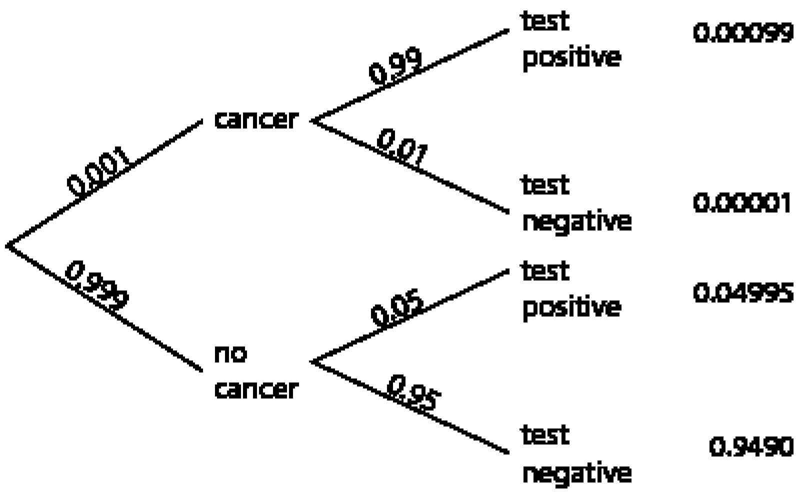


- For Bayesian reasoning, one can expand this structure into a *Bayes tree*:



Final Remarks

- Now we can reason on each conditional probability separately
- E.g. we can update any of the paths based on new experiments
- We can compute probabilities (e.g. the probability of not having cancer) by summing over the respective paths



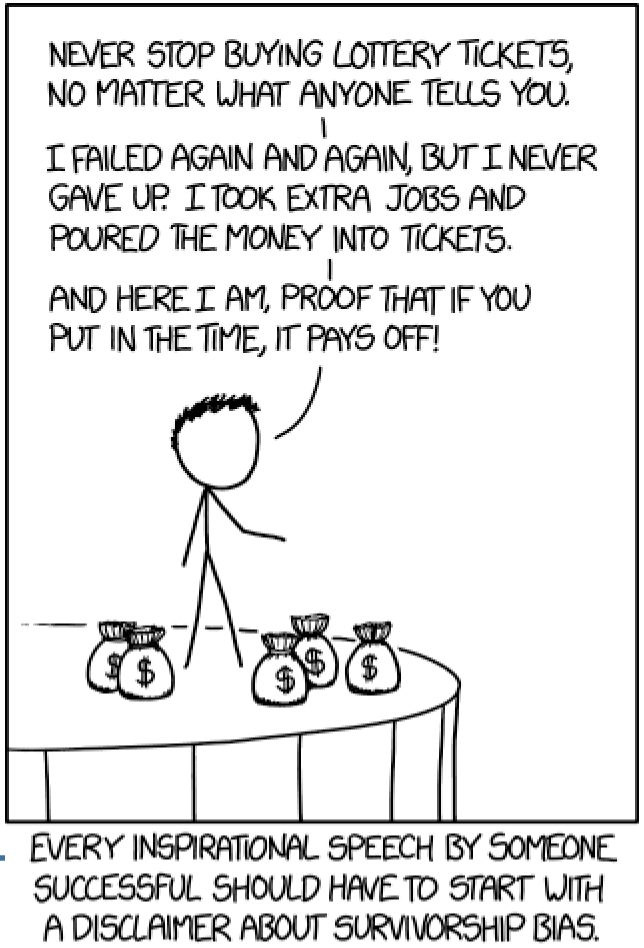
Final Remarks

A final word about probabilities:

- Remember the fundamental concept of frequentist statistics:
  - Repeat an experiment many times
  - Estimate the probability of an event as the fraction of samples in which the event occurred
- But what if we have “experiments” which cannot be repeated?

Final Remarks

- The frequentist philosophy fails in cases where a real-life event fundamentally occurs only once
  - e.g. the outcome of a particular election
- Also a problem how to interpret probabilities *after* an event (image credit: XKCD)
- In complex cases, there are also further practical problems
  - e.g. with frequentist estimation



Final Remarks

- Bayesian statistics consider a “probability” as our *uncertainty* about an event

Final Remarks

- Bayesian statistics consider a “probability” as our *uncertainty* about an event
  - very natural example

will throw a fair die soon
P(1) = … = P(6) = 1/6

→

→

“5”

→

P(5) = 1, all other P(i) = 0

Final Remarks

- Bayesian statistics consider a “probability” as our *uncertainty* about an event
  - very natural example

will throw a fair die soon
P(1) = … = P(6) = 1/6

“5”

P(5) = 1, all other P(i) = 0

  - actually useful example

have only a rough idea about a parameter (high variance)

perform some experiments

have a better estimate

perform many experiments

have a good estimate with which we are satisfied

Frequentist and Bayesian Statistics

	Definition	Frequentist Interpretation	Bayesian Interpretation
P(B)	Probability that event B happens	Assume we repeat an experiment many times. P(B) is the fraction of trials in which B has happened.	P(B) is our "knowledge" of event B at some point.
P(B | A)	Probability that B happens, given that A has "already" happened / is known to have happened	Assume we repeat an experiment many times. P(B|A) is the fraction of trials in which happened B and A, *out of those* where A has happened	P(B | A) is our "knowledge" of B *after A has happened* / after we have observed A

Conclusion / Summary

Istituto Dalle Molle di studi sull'intelligenza artificiale
USI/SUPSI IDSIA

- Of today's lecture, you absolutely should remember Bayes' Theorem, and you should be able to use it
- You will encounter estimation theory in the future
  - …in this lecture and elsewhere
- At least, make sure that you understood the principle of Maximum Likelihood Estimation
- Also remember how to create a classifier from an estimated probability distribution