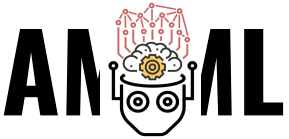
Politecnico di Torino



Advanced Machine Learning

A.A. 2024/2025

Tatiana Tommasi

Basic Statistics - why

machine learning is all about making predictions

diagnosys: predict the probability of a patient suffering a heart attack in the next year, given their clinical history.

anomaly detection: assess how likely a set of readings from an airplane's jet engine would be, when it operates in normal conditions.

reinforcement learning: we want an agent to act intelligently in an environment. This means we need to think about the probability of getting a high reward under each of the available actions.

recommender systems: say hypothetically that we worked for a large online bookseller. We might want to estimate the probability that a particular user would buy a particular book.

Basic Statistics - essential tools for data analysis

The world is a very uncertain place

- Uncertain **inputs**

missing data
noisy data

- Uncertain **knowledge**

Multiple causes lead to multiple effects
Incomplete enumeration of conditions or effects
Incomplete knowledge of causality in the domain
Stochastic effects

- Uncertain **outputs**

induction is inherently uncertain
incomplete deductive inference may be uncertain

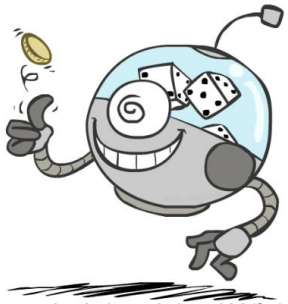
Probability allows to summarize uncertainty from various sources

Sample Spaces

A **sample space** Ω is the set of all possible outcomes of a (conceptual or physical) random experiment. (Ω can be finite or infinite.)

Examples: the set of all possible outcomes of

- Rolling a dice: {1,2,3,4,5,6}
- Flipping a coin: {H, T}
- Flipping a coin three times: {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}
- A person's age: the positive integers
- A person's height: the positive reals



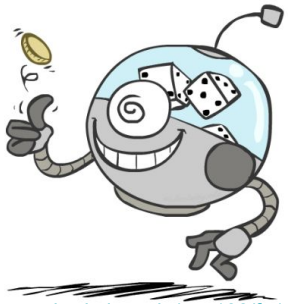
http://www-inst.eecs.berkeley.edu/~cs188/fa19/

Event

An **event** is a subset of the sample space Ω

Examples
- the book is open at an odd number
- rolling a dice: the output number is <4
- a random person's height A : x<A<y

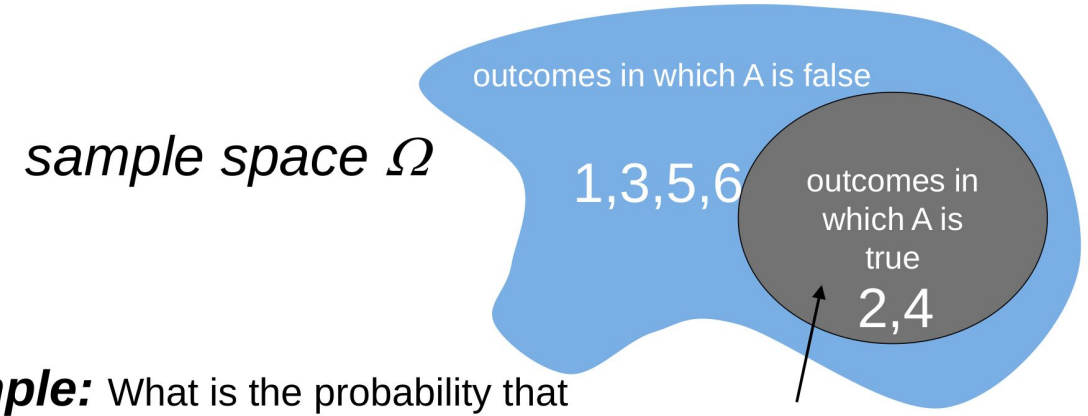
We ask:
What is the probability of a particular event?



http://www-inst.eecs.berkeley.edu/~cs188/fa19/

Probability

Probability P(A): the probability that event (subset) A happens, is a function that maps the event A onto the interval [0, 1] . P(A) is also called the probability measure of A .



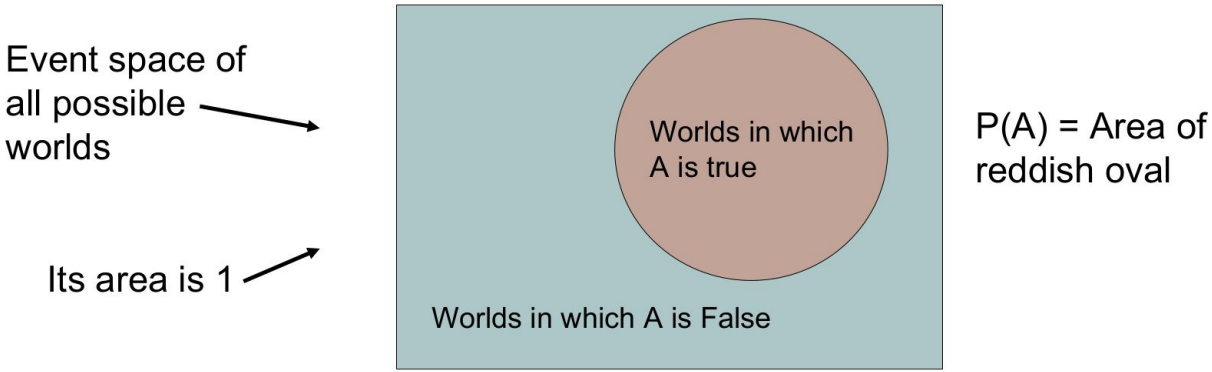
Example: What is the probability that the number on the dice is 2 or 4?

Figure Credit: Barnabás Póczos & Alex Smola

The Axioms of Probability

What defines a reasonable theory of uncertainty?

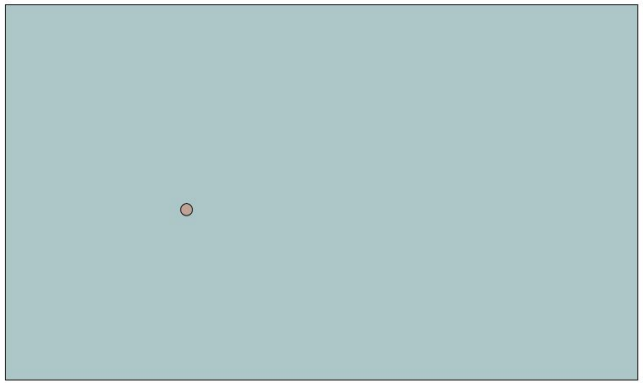
1. All probabilities are between 0 and 1
   $0 \leq P(A) \leq 1$



The Axioms of Probability

What defines a reasonable theory of uncertainty?

2. Valid propositions have probability 1,
   Unsatisfiable propositions have probability 0
   P(empty-set) = 0, P(everything) = 1



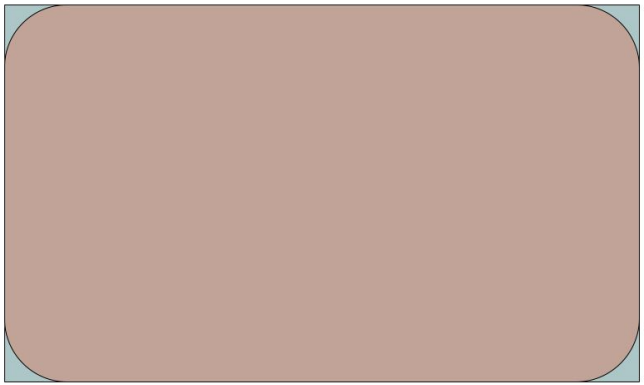
The area of A can t get any smaller than 0

And a zero area would mean no world could ever have A true

The Axioms of Probability

What defines a reasonable theory of uncertainty?

2. Valid propositions have probability 1,
   Unsatisfiable propositions have probability 0
   P(empty-set) = 0, P(everything) = 1



The area of A can t get any bigger than 1

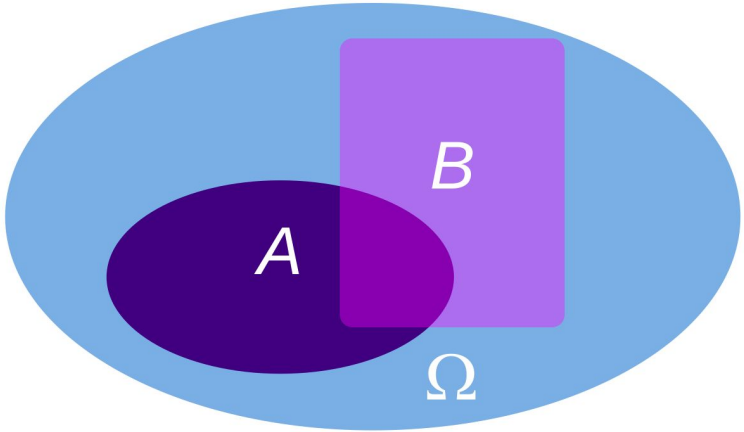
And an area of 1 would mean all worlds will have A true

The Axioms of Probability

What defines a reasonable theory of uncertainty?

3. The probability of a disjunction is given by
   $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Venn Diagram



The Axioms of Probability

What defines a reasonable theory of uncertainty?

1. All probabilities are between 0 and 1
   $0 \leq P(A) \leq 1$

2. Valid propositions have probability 1,
   Unsatisfiable propositions have probability 0
   P(empty-set) = 0, P(everything) = 1

3. The probability of a two events
   $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

The only system with this property:
If you gamble using them you can't be unfairly exploited by an opponent using some other system [De Finetti, 1931]

Random Variables

Real valued **random variable** is a function of the outcome of a randomized experiment

$$X: \Omega \rightarrow \mathbb{R}$$

$$P(a < X < b) \doteq P(\omega : a < X(\omega) < b)$$
$$P(X = a) \doteq P(\omega : X(\omega) = a)$$

Examples
Discrete random variable examples (Ω is discrete):

- X(ω) = True if a randomly drawn person (ω) from our class (Ω) is female
- X(ω) = The hometown X(ω) of a randomly drawn person (ω) from our class (Ω)

Random Variables

Real valued **random variable** is a function of the outcome of a randomized experiment

$$X: \Omega \rightarrow \mathbb{R}$$

$$P(a < X < b) \doteq P(\omega : a < X(\omega) < b)$$
$$P(X = a) \doteq P(\omega : X(\omega) = a)$$

Probabilities from Event

Draw 2 numbers between 1 and 4. Let r.v. X be their sum.																
E	11	12	13	14	21	22	23	24	31	32	33	34	41	42	43	44
X(E)	2	3	4	5	3	4	5	6	4	5	6	7	5	6	7	8

Induced probability function on X.							
x	2	3	4	5	6	7	8
P(X=x)	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

Probability Distribution

Temperature

$P(T)$

T	P
hot	0.5
cold	0.5

Weather

$P(W)$

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

A probability (lower case value) is a single number $P(W = rain) = 0.1$

A distribution is a TABLE of probabilities of values

http://www-inst.eecs.berkeley.edu/~cs188/fa19/

Joint Probability Distribution

Two or more random variables may interact
The probability of one taking on a certain value depends on which value(s) the others are taking.

We call this a joint ensemble and write P(X = x and Y = y)

Temperature

$P(T)$

T	P
hot	0.5
cold	0.5

Weather

$P(W)$

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

http://www-inst.eecs.berkeley.edu/~cs188/fa19/

Event

An event is a set E of outcomes $$P(E) = \sum_{(x_1 \ldots x_n) \in E} P(x_1 \ldots x_n)$$

From a joint distribution, we can calculate the probability of any event

- Probability that it's hot AND sunny?
- Probability that it's hot?
- Probability that it's hot OR sunny?

Typically, the events we care about are partial assignments, like P(T=hot)

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

http://www-inst.eecs.berkeley.edu/~cs188/fa19/

Event

An event is a set E of outcomes $$P(E) = \sum_{(x_1 \ldots x_n) \in E} P(x_1 \ldots x_n)$$

From a joint distribution, we can calculate the probability of any event

- Probability that it's hot AND sunny? (0.4)
- Probability that it's hot? (0.5)
- Probability that it's hot OR sunny? (0.7)

Typically, the events we care about are partial assignments, like P(T=hot)

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

http://www-inst.eecs.berkeley.edu/~cs188/fa19/

Marginal Distribution

- Marginal distributions are sub-tables which eliminate variables
- Marginalization (summing out): Combine collapsed rows by addition

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(t) = \sum_s P(t, s)$$

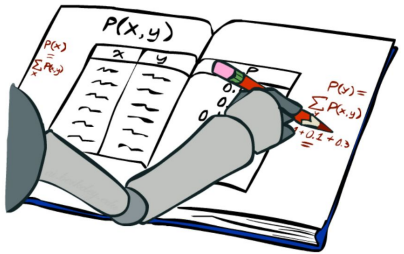
$$P(s) = \sum_t P(t, s)$$

$P(T)$

T	P
hot	0.5
cold	0.5

$P(W)$

W	P
sun	0.6
rain	0.4



$$P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1, X_2 = x_2)$$

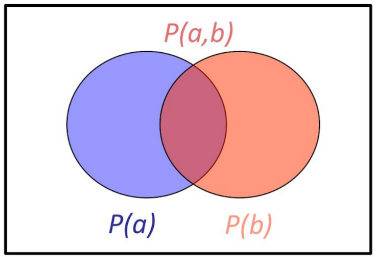
http://www-inst.eecs.berkeley.edu/~cs188/fa19/

Conditional Probabilities

There is a simple relation between joint and marginal probabilities: used to define the conditional probability

P(X|Y) = Fraction of worlds in which X event is true given Y event is true.

$$P(a|b) = \frac{P(a,b)}{P(b)}$$



$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(W = s|T = c) = \frac{P(W = s, T = c)}{P(T = c)} = \frac{0.2}{0.5} = 0.4$$

$$= P(W = s, T = c) + P(W = r, T = c)$$
$$= 0.2 + 0.3 = 0.5$$

http://www-inst.eecs.berkeley.edu/~cs188/fa19/

Conditional Distributions

Conditional distributions are probability distributions over some variables given fixed values of others

$$P(W = s|T = c) = \frac{P(W = s, T = c)}{P(T = c)}$$
$$= \frac{P(W = s, T = c)}{P(W = s, T = c) + P(W = r, T = c)}$$
$$= \frac{0.2}{0.2 + 0.3} = 0.4$$

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$P(W|T = c)$

W	P
sun	0.4
rain	0.6

$$P(W = r|T = c) = \frac{P(W = r, T = c)}{P(T = c)}$$
$$= \frac{P(W = r, T = c)}{P(W = s, T = c) + P(W = r, T = c)}$$
$$= \frac{0.3}{0.2 + 0.3} = 0.6$$

http://www-inst.eecs.berkeley.edu/~cs188/fa19/

Conditional Distributions

Conditional distributions are probability distributions over some variables given fixed values of others

$$P(W = s|T = c) = \frac{P(W = s, T = c)}{P(T = c)}$$
$$= \frac{P(W = s, T = c)}{P(W = s, T = c) + P(W = r, T = c)}$$
$$= \frac{0.2}{0.2 + 0.3} = 0.4$$

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

0.5

SELECT the joint probabilities matching the evidence

$P(c, W)$

T	W	P	
cold	sun	0.2	/ 0.5
cold	rain	0.3	/ 0.5

NORMALIZE the selection (make it sum to one)

$P(W|T = c)$

W	P
sun	0.4
rain	0.6

$$P(W = r|T = c) = \frac{P(W = r, T = c)}{P(T = c)}$$
$$= \frac{P(W = r, T = c)}{P(W = s, T = c) + P(W = r, T = c)}$$
$$= \frac{0.3}{0.2 + 0.3} = 0.6$$

http://www-inst.eecs.berkeley.edu/~cs188/fa19/

Probabilistic Inference

Probabilistic inference: compute a desired probability from other known probabilities (e.g. conditional from joint)

We generally compute conditional probabilities

- P(on time | no reported accidents) = 0.90
- These represent the agent's beliefs given the evidence

Probabilities change with new evidence:

- P(on time | no accidents, 5 a.m.) = 0.95
- P(on time | no accidents, 5 a.m., raining) = 0.80
- Observing new evidence causes beliefs to be updated

http://www-inst.eecs.berkeley.edu/~cs188/fa19/

Inference by Enumeration

- P(W)?

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

http://www-inst.eecs.berkeley.edu/~cs188/fa19/

Inference by Enumeration

- P(W)?

P(sun)=.3+.1+.1+.15=.65

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

http://www-inst.eecs.berkeley.edu/~cs188/fa19/

Inference by Enumeration

- P(W)?

P(sun)=.3+.1+.1+.15=.65

P(rain)=1-.65=.35

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

http://www-inst.eecs.berkeley.edu/~cs188/fa19/

Inference by Enumeration

- P(W | winter, hot)?

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

http://www-inst.eecs.berkeley.edu/~cs188/fa19/

Inference by Enumeration

- P(W | winter, hot)?

P(sun|winter,hot)~.1
P(rain|winter,hot)~.05

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

http://www-inst.eecs.berkeley.edu/~cs188/fa19/

Inference by Enumeration

- P(W | winter, hot)?

P(sun|winter,hot)~.1
P(rain|winter,hot)~.05
P(sun|winter,hot)=2/3
P(rain|winter,hot)=1/3

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

http://www-inst.eecs.berkeley.edu/~cs188/fa19/

Inference by Enumeration

- P(W | winter)?

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

http://www-inst.eecs.berkeley.edu/~cs188/fa19/

Inference by Enumeration

- P(W | winter)?

P(sun|winter)~.25
P(rain|winter)~.25

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

http://www-inst.eecs.berkeley.edu/~cs188/fa19/

Inference by Enumeration

- P(W | winter)?

P(sun|winter)~.25
P(rain|winter)~.25
P(sun|winter)=.5
P(rain|winter)=.5

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

http://www-inst.eecs.berkeley.edu/~cs188/fa19/

The Product & Chain Rule

$$P(y)P(x|y) = P(x,y)$$

$P(W)$

R	P
sun	0.8
rain	0.2

$P(D|W)$

D	W	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3

http://www-inst.eecs.berkeley.edu/~cs188/fa19/

The Product & Chain Rule

$$P(y)P(x|y) = P(x,y)$$

$P(W)$

R	P
sun	0.8
rain	0.2

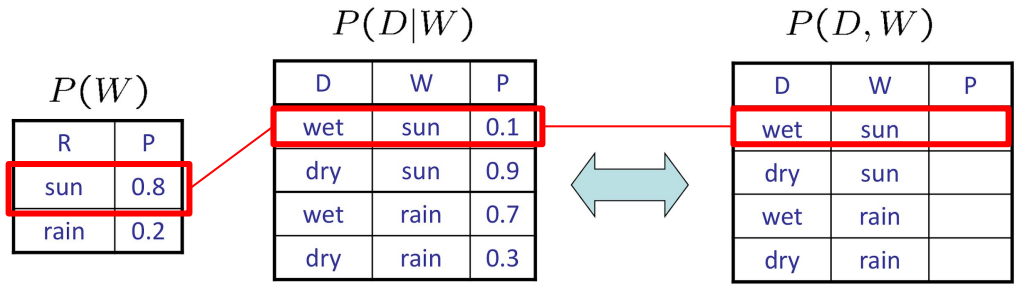
$P(D|W)$

D	W	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3

http://www-inst.eecs.berkeley.edu/~cs188/fa19/

The Product & Chain Rule

$$P(y)P(x|y) = P(x,y)$$



http://www-inst.eecs.berkeley.edu/~cs188/fa19/

The Product & Chain Rule

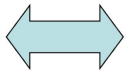
$$P(y)P(x|y) = P(x,y)$$

$P(W)$

R	P
sun	0.8
rain	0.2

$P(D|W)$

D	W	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3



$P(D,W)$

D	W	P
wet	sun	
dry	sun	
wet	rain	
dry	rain	

We can always write any joint distribution as an incremental product of conditional distributions

$$P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)$$

$$P(x_1, x_2, \ldots x_n) = \prod_i P(x_i|x_1 \ldots x_{i-1})$$

$$\Pr(X,Y,Z) = \frac{\Pr(X,Y,Z)}{P(Y,Z)} \frac{P(Y,Z)}{P(Z)} P(Z)$$

http://www-inst.eecs.berkeley.edu/~cs188/fa19/

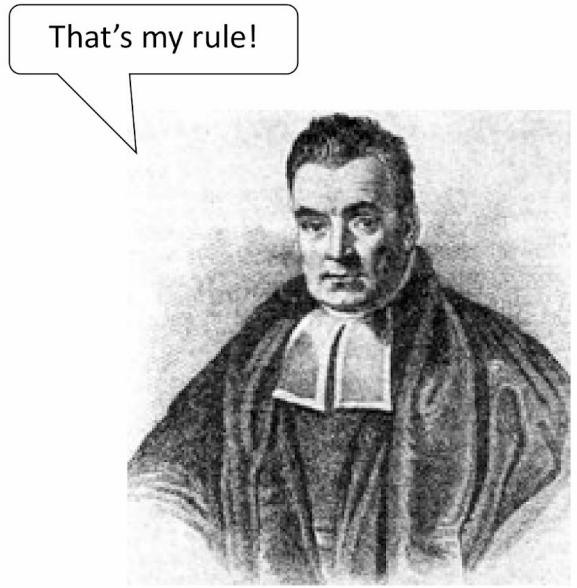
Bayes' Rule

- Two ways to factor a joint distribution over two variables:

$$P(x, y) = P(x|y)P(y) = P(y|x)P(x)$$

- Dividing, we get:

$$P(x|y) = \frac{P(y|x)}{P(y)}P(x)$$



http://www-inst.eecs.berkeley.edu/~cs188/fa19/

Bayes' Rule

- Two ways to factor a joint distribution over two variables:

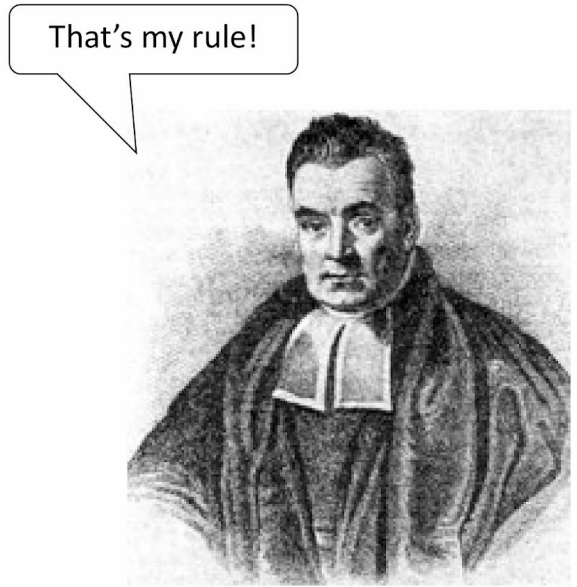
$$P(x, y) = P(x|y)P(y) = P(y|x)P(x)$$

- Dividing, we get:

$$P(x|y) = \frac{P(y|x)}{P(y)}P(x)$$

- Why is this at all helpful?
  - Lets us build one conditional from its reverse
  - Often one conditional is tricky but the other one is simple
  - Foundation of many systems

- In the running for most important AI equation!



http://www-inst.eecs.berkeley.edu/~cs188/fa19/

Bayes' Rule

- Two ways to factor a joint distribution over two variables:

$$P(x, y) = P(x|y)P(y) = P(y|x)P(x)$$

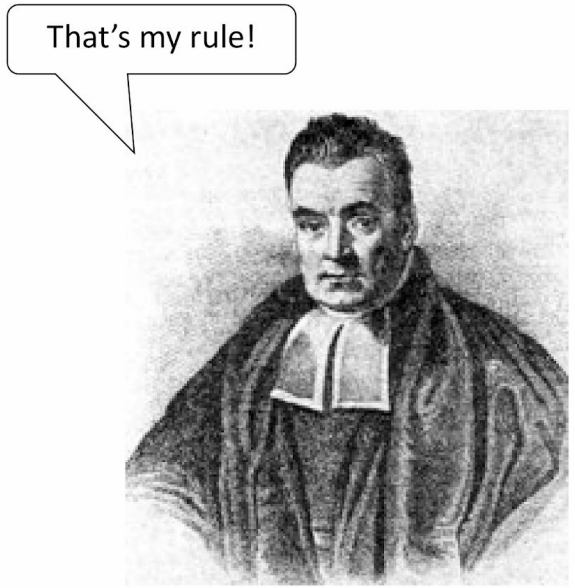
- Dividing, we get:

posterior $P(x|y) \propto \frac{P(y|x)}{P(y)} P(x)$ prior

likelihood

- Why is this at all helpful?
  - Lets us build one conditional from its reverse
  - Often one conditional is tricky but the other one is simple
  - Foundation of many systems

- In the running for most important AI equation!



http://www-inst.eecs.berkeley.edu/~cs188/fa19/

Bayes' Rule

- Allows us to reason from **evidence** to **hypotheses**
- Another way of thinking about Bayes' rule:

$$P(\text{hypothesis} \mid \text{evidence}) = \frac{P(\text{evidence} \mid \text{hypothesis}) \times P(\text{hypothesis})}{P(\text{evidence})}$$

In the flu example:

$\mathrm{P}(\text{headache}) = 1/10$ $\mathrm{P}(\text{flu}) = 1/40$

$\mathrm{P}(\text{headache} \mid \text{flu}) = 1/2$

Given evidence of headache, what is $\mathrm{P}(\text{flu} \mid \text{headache})$?

Solve via Bayes rule!

Slide Credit: Eric Eaton

Bayes' Rule

Example: Diagnostic probability from causal probability

$$P(\text{cause}|\text{effect}) = \frac{P(\text{effect}|\text{cause})P(\text{cause})}{P(\text{effect})}$$

- M: meningitis, S: stiff neck

$$\left.\begin{aligned} P(+m) &= 0.0001 \\ P(+s|+m) &= 0.8 \\ P(+s|-m) &= 0.01 \end{aligned}\right\}$$ Example givens

$$P(+m|+s) = \frac{P(+s|+m)P(+m)}{P(+s)} = \frac{P(+s|+m)P(+m)}{P(+s|+m)P(+m) + P(+s|-m)P(-m)} = \frac{0.8 \times 0.0001}{0.8 \times 0.0001 + 0.01 \times 0.999}$$

- Note: posterior probability of meningitis still very small

http://www-inst.eecs.berkeley.edu/~cs188/fa20/

The Monty Hall Game Show Problem

https://www.youtube.com/watch?v=cXqDIFUB7YU

The Monty Hall Game Show Problem

Question: In a TV Game show, a contestant selects one of three doors; behind one of the doors there is a prize, and behind the other two there are no prizes. After the contestant selects a door, the game-show host opens one of the remaining doors, and reveals that there is no prize behind it. The host then asks the contestant whether they want to SWITCH their choice to the other unopened door, or STICK to their original choice.

Is it probabilistically advantageous for the contestant to SWITCH doors, or is the probability of winning the prize the same whether they STICK or SWITCH? (Assume that the host selects a door to open, from those available, with equal probability).

The Monty Hall Game Show Problem

Question: In a TV Game show, a contestant selects one of three doors; behind one of the doors there is a prize, and behind the other two there are no prizes. After the contestant selects a door, the game-show host opens one of the remaining doors, and reveals that there is no prize behind it. The host then asks the contestant whether they want to SWITCH their choice to the other unopened door, or STICK to their original choice.

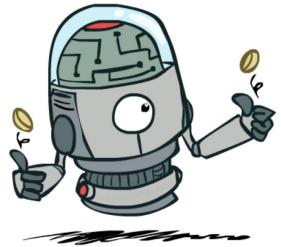
Is it probabilistically advantageous for the contestant to SWITCH doors, or is the probability of winning the prize the same whether they STICK or SWITCH? (Assume that the host selects a door to open, from those available, with equal probability).

A = prize behind door A - selected by the contestant
B = prize behind door B - remained closed
C = prize behind door C - opened by the host
H_C = the host opens door C

$P(A|H_C)$ = STICK
$P(B|H_C)$ = SWITCH

Independence

Y and X don't contain information about each other.
Observing Y doesn't help predicting X.
Observing X doesn't help predicting Y.



http://www-inst.eecs.berkeley.edu/~cs188/fa20/

Independent random variables:

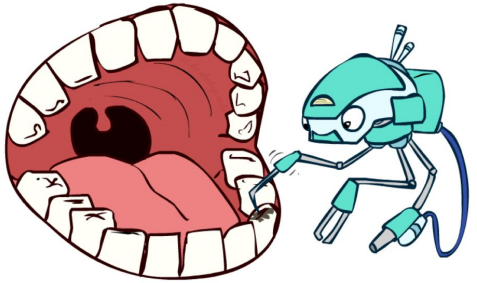
$$P(X,Y) = P(X)P(Y)$$

$$P(X|Y) = P(X)$$

Examples

Independent: Winning on roulette this week and next week.
Dependent: Russian roulette

Conditionally Independent



P(Toothache, Cavity, Catch)

If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
P(+catch | +toothache, +cavity) = P(+catch | +cavity)

The same independence holds if I don't have a cavity:
P(+catch | +toothache,-cavity) = P(+catch|-cavity)

Catch is conditionally independent of Toothache given Cavity:
P(Catch | Toothache, Cavity) = P(Catch | Cavity)

Equivalent statements:
P(Toothache | Catch , Cavity) = P(Toothache | Cavity)
P(Toothache, Catch | Cavity) = P(Toothache | Cavity) P(Catch | Cavity)

http://www-inst.eecs.berkeley.edu/~cs188/fa20/

Conditionally Independent

Conditionally independent:

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

Knowing Z makes X and Y independent

London taxi drivers: A survey has pointed out a positive and significant correlation between the **number of accidents** and wearing **coats**. They concluded that coats could hinder movements of drivers and be the cause of accidents. A new law was prepared to prohibit drivers from wearing coats when driving.

Finally another study pointed out that people wear coats when it rains...

$$P(\text{Accidents}, \text{Coats} \mid \text{Rain}) = P(\text{Accidents} \mid \text{Rain})P(\text{Coats} \mid \text{Rain})$$

Equivalent to:

$$(\forall x, y, z)P(X = x|Y = y, Z = z) = P(X = x|Z = z)$$

Model Based Classification with Naïve Bayes

Machine learning: how to acquire a model from data / experience

$y = f(x) \rightarrow$ Learning parameters (e.g. probabilities)

Model Based Classification with Naïve Bayes

Machine learning: how to acquire a model from data / experience

$y = f(x) \rightarrow$ Learning parameters (e.g. probabilities)

Task: Predict whether or not a picnic spot is enjoyable

Training Data: $X = (X_1 \quad X_2 \quad X_3 \quad \ldots \quad \ldots \quad X_d) \quad Y$

n rows

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

Model Based Classification with Naïve Bayes

$$P(y|\mathbf{x}) = \frac{P(\mathbf{x}|y)P(y)}{P(\mathbf{x})}$$

Estimating $P(y)$ is generally easy.
Estimating $P(\mathbf{x}|y)$, however, is not easy!

Make it easier: assume that the attributes are independent given the class label

$$P(X_1 ... X_d|Y) = \prod_{i=1}^{d} P(X_i|Y)$$

Task: Predict whether or not a picnic spot is enjoyable

Training Data: $\mathbf{X} = (X_1 \quad X_2 \quad X_3 \quad \ldots \quad \ldots \quad X_d) \quad Y$

n rows

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

Slide Credit: Barnabás Póczos & Alex Smola

Model Based Classification with Naïve Bayes

How many parameters to estimate?

(X is composed of d binary features,
Y has K possible class labels)

~2^dK vs ~2dK

$$P(X_1 \ldots X_d|Y) = \prod_{i=1}^{d} P(X_i|Y)$$

Task: Predict whether or not a picnic spot is enjoyable

Training Data: **X = (X_1 X_2 X_3 ... ... X_d) Y**

n rows

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

Slide Credit: Barnabás Póczos & Alex Smola

Naïve Bayes Classifier

Given:

- Class prior $P(Y)$
- d conditionally independent features $X_1,\ldots X_d$ given the class label Y
- For each X_i feature, we have the conditional likelihood $P(X_i|Y)$

- Can produce probabilities by:
  - For each possible class label y_k, compute

$$\tilde{P}(Y = y_k \mid X = \mathbf{x}) = P(Y = y_k) \prod_{j=1}^{d} P(X_j = x_j \mid Y = y_k)$$

Naïve Bayes Classifier

Given:

- Class prior $P(Y)$
- d conditionally independent features $X_1, \ldots X_d$ given the class label Y
- For each X_i feature, we have the conditional likelihood $P(X_i|Y)$

$$P(y|\mathbf{x}) = \frac{P(\mathbf{x}|y)P(y)}{P(\mathbf{x})}$$

predictions

- Can produce probabilities by:
  - For each possible class label y_k, compute

$$\underbrace{\tilde{P}(Y = y_k \mid X = \mathbf{x})}_{\text{This is the numerator of Bayes rule, and is therefore off the true probability by a factor that makes probabilities sum to 1}} = P(Y = y_k) \prod_{j=1}^{d} P(X_j = x_j \mid Y = y_k)$$

Naïve Bayes Classifier

Given:

- Class prior $P(Y)$
- d conditionally independent features $X_1,\ldots X_d$ given the class label Y
- For each X_i feature, we have the conditional likelihood $P(X_i|Y)$

Naïve Bayes Decision rule:

$$f_{NB}(\mathbf{x}) = \arg\max_y P(x_1,\ldots,x_d \mid y)P(y)$$

$$= \arg\max_y \prod_{i=1}^{d} P(x_i|y)P(y)$$

Slide Credit: Barnabás Póczos & Alex Smola

Naïve Bayes Classifier Example

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Slide Credit: Ana Teresa Freitas

Naïve Bayes Classifier Example

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

- Learning Phase

Outlook	Play=*Yes*	Play=*No*
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temperature	Play=*Yes*	Play=*No*
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Play=*Yes*	Play=No
High	3/9	4/5
Normal	6/9	1/5

Wind	Play=*Yes*	Play=*No*
Strong	3/9	3/5
Weak	6/9	2/5

$P(\text{Play}=Yes) = 9/14$

$P(\text{Play}=No) = 5/14$

Slide Credit: Ana Teresa Freitas

Naïve Bayes Classifier Example

- Test Phase
  - Given a new instance, predict its label

    $\mathbf{x}'$=(Outlook=*Sunny*, Temperature=*Cool*, Humidity=*High*, Wind=*Strong*)
  - Look up tables achieved in the learning phrase

P(Outlook=*Sunny* | Play=*Yes*) = 2/9
P(Temperature=*Cool* | Play=*Yes*) = 3/9
P(Huminity=*High* | Play=*Yes*) = 3/9
P(Wind=*Strong* | Play=*Yes*) = 3/9
P(Play=*Yes*) = 9/14

P(Outlook=*Sunny* | Play=*No*) = 3/5
P(Temperature=*Cool* | Play==*No*) = 1/5
P(Huminity=*High* | Play=*No*) = 4/5
P(Wind=*Strong* | Play=*No*) = 3/5
P(Play=*No*) = 5/14

Slide Credit: Ana Teresa Freitas

Naïve Bayes Classifier Example

- Test Phase
  - Given a new instance, predict its label

    $\mathbf{x}'$=(Outlook=*Sunny*, Temperature=*Cool*, Humidity=*High*, Wind=*Strong*)
  - Look up tables achieved in the learning phrase

    P(Outlook=*Sunny* | Play=*Yes*) = 2/9
    P(Temperature=*Cool* | Play=*Yes*) = 3/9
    P(Huminity=*High* | Play=*Yes*) = 3/9
    P(Wind=*Strong* | Play=*Yes*) = 3/9
    P(Play=*Yes*) = 9/14

    P(Outlook=*Sunny* | Play=*No*) = 3/5
    P(Temperature=*Cool* | Play==*No*) = 1/5
    P(Huminity=*High* | Play=*No*) = 4/5
    P(Wind=*Strong* | Play=*No*) = 3/5
    P(Play=*No*) = 5/14
  - Decision making

    $P(Yes \mid \mathbf{x}') \approx [P(Sunny \mid Yes)P(Cool \mid Yes)P(High \mid Yes)P(Strong \mid Yes)]P(Play=Yes) = 0.0053$

    $P(No \mid \mathbf{x}') \approx [P(Sunny \mid No)\ P(Cool \mid No)P(High \mid No)P(Strong \mid No)]P(Play=No) = 0.0206$

    Given the fact $P(Yes \mid \mathbf{x}') < P(No \mid \mathbf{x}')$, we label $\mathbf{x}'$ to be "*No*".

Slide Credit: Ana Teresa Freitas

Naïve Bayes Classifier Summary

- Computationally very fast
  - Training: only one pass over the training set
  - Classification: linear in the number of attributes (features)

- Despite its conditional independence assumption, Naïve Bayes classifier shows a good performance in several application domains

When to use?

- A moderate or large training set available with instances represented by a large number of attributes

Laplace Smoothing

Issue: It might happen that an attribute never appears for a specific class

Sky	Temp	Humid	Wind	Water	Forecast	Play?
sunny	warm	normal	strong	warm	same	*yes*
sunny	warm	high	strong	warm	same	*yes*
rainy	cold	high	strong	warm	change	*no*
sunny	warm	high	strong	cool	change	*yes*

$P(play) = 3/4$ $\quad P(\neg play) = 1/4$

$P(Sky = sunny \mid play) = 1$ $\quad P(Sky = sunny \mid \neg play) = 0$

$P(Humid = high \mid play) = 2/3$ $\quad P(Humid = high \mid \neg play) = 1$

... $\quad$ ...

Laplace Smoothing

Fix by using Laplace smoothing:

- Adds 1 to each count

$$P(X_j = v \mid Y = y_k) = \frac{\mathrm{c}_v + 1}{\sum_{v' \in \mathrm{values}(X_j)} \mathrm{c}_{v'} + |\mathrm{values}(X_j)|}$$

where

- c_v is the count of training instances with a value of v for attribute j and class label y_k
- $|\mathrm{values}(X_j)|$ is the number of values X_j can take on

Laplace Smoothing

Sky	Temp	Humid	Wind	Water	Forecast	Play?
sunny	warm	normal	strong	warm	same	*yes*
sunny	warm	high	strong	warm	same	*yes*
rainy	cold	high	strong	warm	change	*no*
sunny	warm	high	strong	cool	change	*yes*

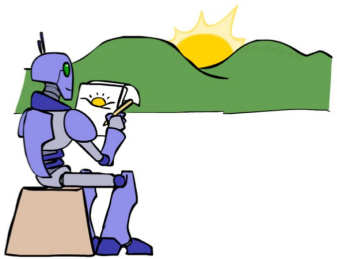
$P(play) = 3/4$ $\quad\quad$ $P(\neg play) = 1/4$

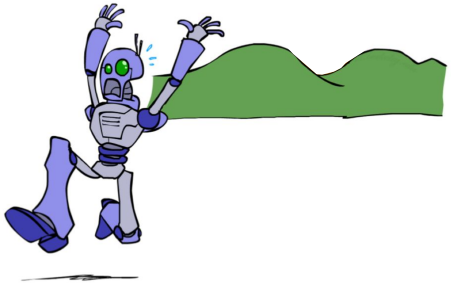
$P(Sky = sunny \mid play) = 4/5$ $\quad\quad$ $P(Sky = sunny \mid \neg play) = 1/3$

$P(Humid = high \mid play) = 3/5$ $\quad\quad$ $P(Humid = high \mid \neg play) = 2/3$

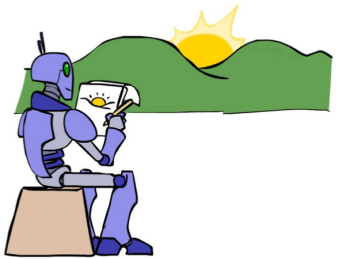
... $\quad\quad$ ...

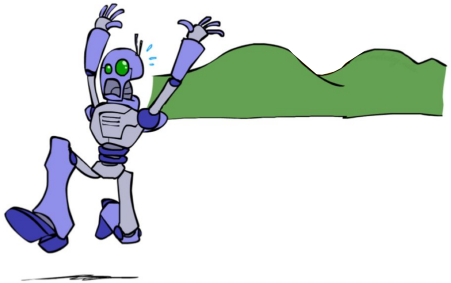
Laplace Smoothing





Laplace Smoothing

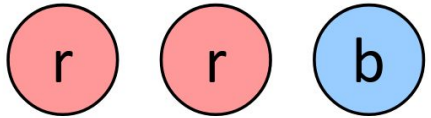




- Laplace's estimate (extended):
  - Pretend you saw every outcome k extra times

$$P_{LAP,k}(x) = \frac{c(x) + k}{N + k|X|}$$

  - What's Laplace with k = 0?
  - k is the strength of the prior



$$P_{LAP,0}(X) = \left\langle \frac{2}{3}, \frac{1}{3} \right\rangle$$

$$P_{LAP,1}(X) = \left\langle \frac{3}{5}, \frac{2}{5} \right\rangle$$

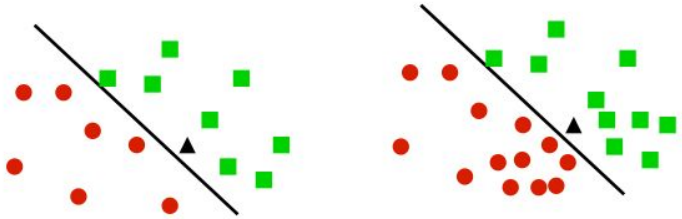
$$P_{LAP,100}(X) = \left\langle \frac{102}{203}, \frac{101}{203} \right\rangle$$

Discriminative vs Generative Learning

Many supervised learning can be viewed as estimating $P(X,Y)$. Generally thy fall into two categories

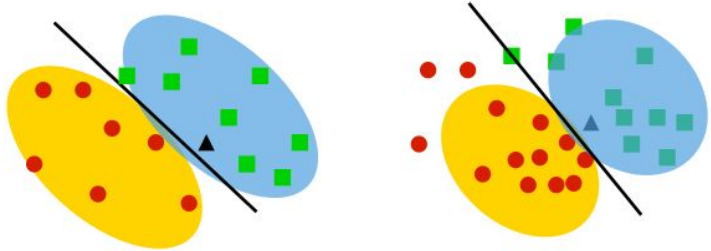
- When we estimate $P(X,Y)=P(X|Y)P(Y)$ then we call it *generative learning.*
- When we only estimate $P(Y|X)$ then we call it *discriminative learning.*

Discriminative



- Only care about estimating the conditional probabilities
- Very good when underlying distribution of data is really complicated (e.g. texts, images, movies)

Generative



- Model observations (x,y) first
- Then infer p(y|x)
- Good for missing variables, better diagnostics
- Easy to add prior knowledge about data

Slide Credit: Barnabás Póczos & Alex Smola

Summary

- Basic Probability
- Bayes' Rule
- Naive Bayes classifier
- Discriminative vs Generative Learning

The Monty Hall Game Show Problem: Switch!

Question: In a TV Game show, a contestant selects one of three doors; behind one of the doors there is a prize, and behind the other two there are no prizes. After the contestant selects a door, the game-show host opens one of the remaining doors, and reveals that there is no prize behind it. The host then asks the contestant whether they want to SWITCH their choice to the other unopened door, or STICK to their original choice.

Is it probabilistically advantageous for the contestant to SWITCH doors, or is the probability of winning the prize the same whether they STICK or SWITCH? (Assume that the host selects a door to open, from those available, with equal probability).

A = prize behind door A - selected by the contestant
B = prize behind door B - remained closed
C = prize behind door C - opened by the host
H_C = the host opens door C

$P(A|H_C)$ = STICK
$P(B|H_C)$ = SWITCH

Originally $P(A)=P(B)=P(C)=⅓$

$P(H_C|A)=½$; $P(H_C|B)=1$; $P(H_C|C)=0$

$$P(B|H_C)=\frac{P(H_C|B)P(B)}{P(H_C|B)P(B)+P(H_C|A)P(A)+P(H_C|C)P(C)}$$

$$= (⅓)/(1*⅓ + ½*⅓ + 0* ⅓) = ⅔$$

$P(A|H_C)= ⅓$

Maximum a Posteriori (MAP) & Bayesian Learning

- hypothesis **h**: probabilistic theory about how the domain works
- bayesian learning: calculates the probability of each hypothesis and, given the data, makes predictions on that basis.
  →predictions are made by using **all** the hypotheses weighted by their probabilities

d = observed data
X = unknown quantity

$$\mathbf{P}(X \mid \mathbf{d}) = \sum_i \mathbf{P}(X \mid h_i) P(h_i \mid \mathbf{d})$$

Predictions made according to an MAP hypothesis h_{MAP} are approximately Bayesian to the extent that

$$\mathbf{P}(X \mid \mathbf{d}) \approx \mathbf{P}(X \mid h_{\mathrm{MAP}})$$

Maximum a Posteriori (MAP) & Bayesian Learning

$P(y|\mathbf{x}) \propto P(\mathbf{x}|y)P(y)$

posterior likelihood prior

Naïve Bayes Decision rule:

$$f_{NB}(\mathbf{x}) = \arg\max_y P(x_1, \ldots, x_d \mid y)P(y)$$

$$= \arg\max_y \prod_{i=1}^{d} P(x_i|y)P(y)$$

$$P(h_i \mid \mathbf{d}) = \alpha P(\mathbf{d} \mid h_i)P(h_i) \qquad P(\mathbf{d} \mid h_i) = \prod_j P(d_j \mid h_i)$$

$$\mathbf{P}(X \mid \mathbf{d}) = \sum_i \mathbf{P}(X \mid h_i)P(h_i \mid \mathbf{d}) \qquad \mathbf{P}(X \mid \mathbf{d}) \approx \mathbf{P}(X \mid h_{\text{MAP}})$$

when **conditional independence** is satisfied, Naive Bayes corresponds to MAP classification.

Independent and Identically distributed (i.i.d)

- **stationary assumption**: there is a probability distribution over examples that remains stationary over time
- each example data point (before we see it) is a random variable Ej whose observed value ej = (xj, yj) is sampled from that distribution and is independent of the previous examples

$$\mathbf{P}(E_j | E_{j-1}, E_{j-2}, \ldots) = \mathbf{P}(E_j)$$

$$\mathbf{P}(E_j) = \mathbf{P}(E_{j-1}) = \mathbf{P}(E_{j-2}) = \cdots$$

i.i.d: independent and identically distributed

The iid assumption connects the past to the future, without some such connections, all the bets are off and the future could be anything

Maximum A Posteriori → Maximum Likelihood (ML)

- the prior can be used to **penalize complexity**
- more complex hypotheses have lower prior probability

$$\underbrace{P(y|\mathbf{x})}_{\text{posterior}} \propto \underbrace{P(\mathbf{x}|y)}_{\text{likelihood}} \underbrace{P(y)}_{\text{prior}}$$

- if we consider a **uniform prior over the space of hypotheses**, the posterior is maximized by the same hypothesis that maximizes the likelihood→Maximum Likelihood (ML)

- it provides a good approximation to Bayesian and MAP learning when the data set is large

Estimating Probabilities

We are often faced with the situation of having random data which we know (or believe) is drawn from a parametric model, whose parameters we do not know.

Example: the outcome of tossing a coin described by a distribution with unknown parameter θ.

Data, *D* =

$$D = \{X_i\}_{i=1}^{n},\ X_i \in \{\mathrm{H}, \mathrm{T}\}$$

P(Heads) = θ, P(Tails) = 1-θ

In this case we would like to use the data to estimate the value of the parameter θ.

Flips are i.i.d.:

- Independent events
- Identically distributed

Maximum Likelihood Estimation: choose θ that maximizes the probability of observed data

Maximum Likelihood Estimation

Data, D = 

$$D = \{X_i\}_{i=1}^{n},\ X_i \in \{\mathrm{H}, \mathrm{T}\}$$

P(Heads) = θ, P(Tails) = 1-θ

$$\begin{aligned}
\hat{\theta}_{MLE} &= \arg\max_{\theta} P(D \mid \theta) \\
&= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta) \qquad \text{Independent draws} \\
&= \arg\max_{\theta} \prod_{i:X_i=H} \theta \prod_{i:X_i=T} (1-\theta) \qquad \text{Identically distributed} \\
&= \arg\max_{\theta} \underbrace{\theta^{\alpha_H}(1-\theta)^{\alpha_T}}_{J(\theta)}
\end{aligned}$$

$$\frac{\partial J(\theta)}{\partial \theta} = \alpha_H \theta^{\alpha_H - 1}(1-\theta)^{\alpha_T} - \alpha_T \theta^{\alpha_H}(1-\theta)^{\alpha_T - 1}\Big|_{\theta=\hat{\theta}_{\mathrm{MLE}}} = 0$$

$$\alpha_H(1-\theta) - \alpha_T\theta\Big|_{\theta=\hat{\theta}_{\mathrm{MLE}}} = 0$$

$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

Maximum Likelihood Estimation

Data, D = 

$$D = \{X_i\}_{i=1}^{n},\ X_i \in \{\mathrm{H}, \mathrm{T}\}$$

P(Heads) = θ, P(Tails) = 1-θ

I have a coin, I flip it, what is the probability of getting head?
The estimated probability is **%** = frequency of heads

Why? Frequency of heads is exactly the MLE for this problem

$$\frac{\partial J(\theta)}{\partial \theta} = \alpha_H \theta^{\alpha_H - 1}(1-\theta)^{\alpha_T} - \alpha_T \theta^{\alpha_H}(1-\theta)^{\alpha_T - 1}\Big|_{\theta=\hat{\theta}_{\mathrm{MLE}}} = 0$$

$$\alpha_H(1-\theta) - \alpha_T \theta\Big|_{\theta=\hat{\theta}_{\mathrm{MLE}}} = 0$$

$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

slide credit: Barnabás Póczos & Alex Smola

Maximum Likelihood Estimation

Data, D = 

$$D = \{X_i\}_{i=1}^{n},\ X_i \in \{\mathrm{H}, \mathrm{T}\}$$

P(Heads) = θ, P(Tails) = 1-θ

I have a coin, I flip it, what is the probability of getting head?
The estimated probability is **%** = frequency of heads

Why? Frequency of heads is exactly the MLE for this problem

Laid out one standard method for parameter learning:

1. write down an expression for likelihood of the data as a function of the parameter(s)
2. write down the derivative with respect to each parameter
3. find the parameter(s) values such that the derivatives are zero

Discrete Probability Distributions

Bernoulli Distribution -- Ber(θ)
Single toss of a (possibly biased) coin

Ω = {head, tail}
P(head) = θ, P(tails) = 1-θ

$$\mathcal{X} = \{0, 1\}$$
$$0 \leq \theta \leq 1$$
$$\text{Ber}(x \mid \theta) = \theta^{\delta(x,1)}(1-\theta)^{\delta(x,0)}$$

Discrete Probability Distributions

Bernoulli Distribution -- Ber(θ)
Single toss of a (possibly biased) coin

Ω = {head, tail}
P(head) = θ, P(tails) = 1-θ

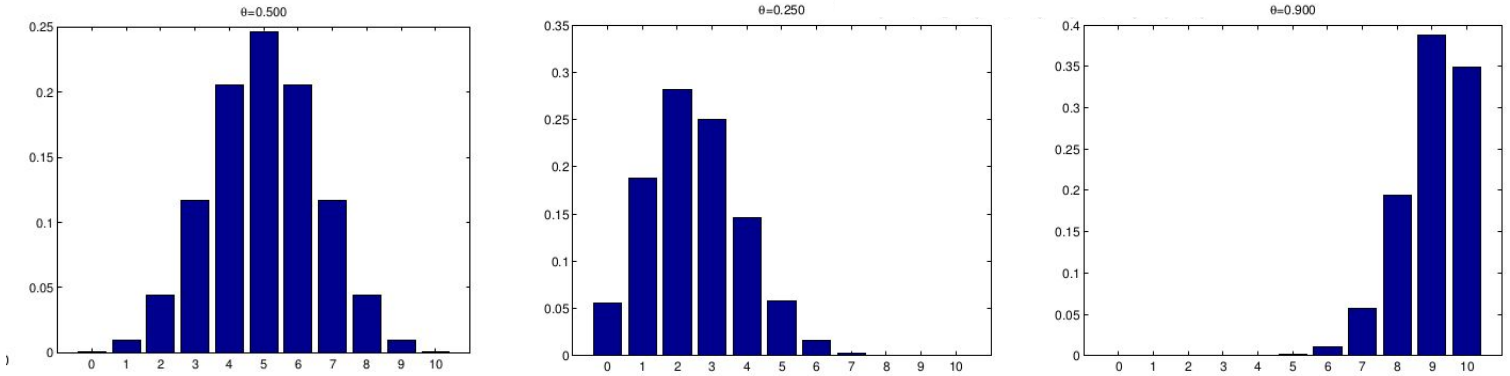
$$\mathcal{X} = \{0, 1\}$$
$$0 \leq \theta \leq 1$$
$$\text{Ber}(x \mid \theta) = \theta^{\delta(x,1)}(1-\theta)^{\delta(x,0)}$$

Binomial Distribution -- Bin(n,θ)
Toss a single (possibly biased) coin n times and report the number k of times it comes up head and n-k it comes up tail.

Ω = {n-long head/tail series} , $|\Omega|=2^n$
number of heads k={0,1,2,...,n}

$$0 \leq \theta \leq 1$$
$$\text{Bin}(k \mid n, \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$
$$\binom{n}{k} = \frac{n!}{(n-k)!k!}$$

Discrete Probability Distributions



Binomial Distribution -- Bin(n,θ)

Toss a single (possibly biased) coin n times and report the number k of times it comes up head and n-k it comes up tail.

Ω = {n-long head/tail series} , $|\Omega|=2^n$

number of heads k={0,1,2,...,n}

$$0 \leq \theta \leq 1$$

$$\mathrm{Bin}(k \mid n, \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}$$

Maximum a Posteriori for coin flip

$$P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta) P(\theta)$$

posterior likelihood prior

Likelihood function is binomial

$$P(\mathcal{D} \mid \theta) = \binom{n}{\alpha_H} \theta^{\alpha_H} (1-\theta)^{\alpha_T}$$

Beta Prior distribution

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

slide credit: Carlos Guestrin

Maximum a Posteriori for coin flip

$$P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta) P(\theta)$$

posterior likelihood prior

Likelihood function is binomial

$$P(\mathcal{D} \mid \theta) = \binom{n}{\alpha_H} \theta^{\alpha_H} (1-\theta)^{\alpha_T}$$

Beta Prior distribution

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

⇒ posterior is Beta distribution

$$P(\theta \mid \mathcal{D}) = \frac{\theta^{\beta_H + \alpha_H - 1}(1-\theta)^{\beta_T + \alpha_T - 1}}{B(\beta_H + \alpha_H, \beta_T + \alpha_T)}$$

$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

P(θ) and P(θ|D) have the same form! [Conjugate prior]

Maximum a Posteriori for coin flip

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta|D)$$

$$= \frac{\beta_H + \alpha_H - 1}{(\beta_H + \alpha_H - 1) + (\beta_T + \alpha_T - 1)}$$

Beta prior is equivalent to extra flips: when the number of flips is high it is 'forgotten' but when the number of flips is small prior is important!

⇒ posterior is Beta distribution

$$P(\theta \mid \mathcal{D}) = \frac{\theta^{\beta_H + \alpha_H - 1}(1-\theta)^{\beta_T + \alpha_T - 1}}{B(\beta_H + \alpha_H, \beta_T + \alpha_T)}$$

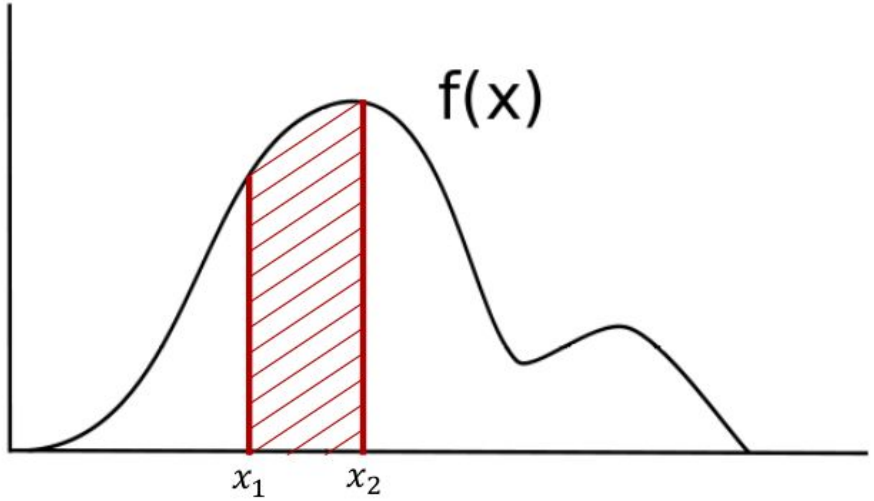
$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

P(θ) and P(θ|D) have the same form! [Conjugate prior]

Continuous Random Variables

A random variable X is **continuous** if its set of possible values is an entire interval of numbers

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f_x(x)dx$$



This explains why P(X=x)=0 for continuous distributions

$$P(X = x) \leq \lim_{\substack{\epsilon \to 0 \\ \epsilon > 0}} [F_x(x) - F_x(x - \epsilon)] = 0$$

Probability Density Function

Pdf properties:

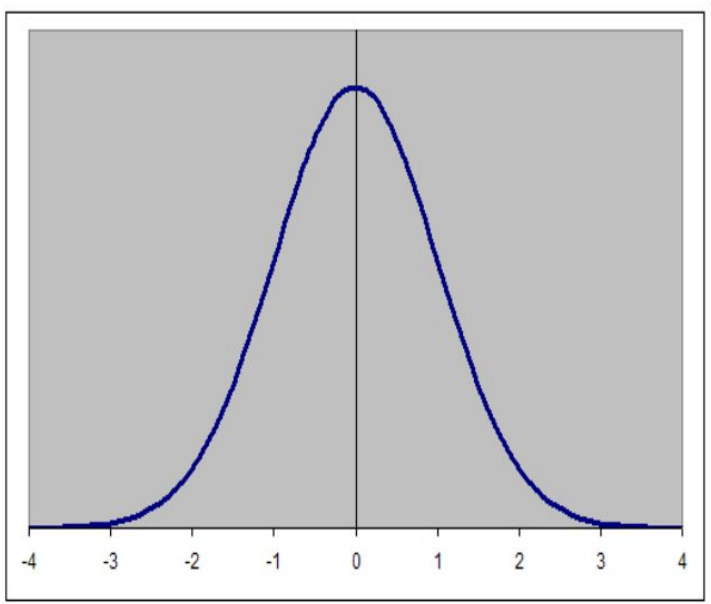
$$f(x) \geq 0$$

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$f(x) = \frac{d}{dx}F(x)$$

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx$$



Intuitively, one can think of f(x)dx as being the probability of X falling within the infinitesimal interval [x, x+dx]. P(x < X < x+dx) = f(x) dx

Probability Density Function

Pdf properties:

$$f(x) \geq 0$$

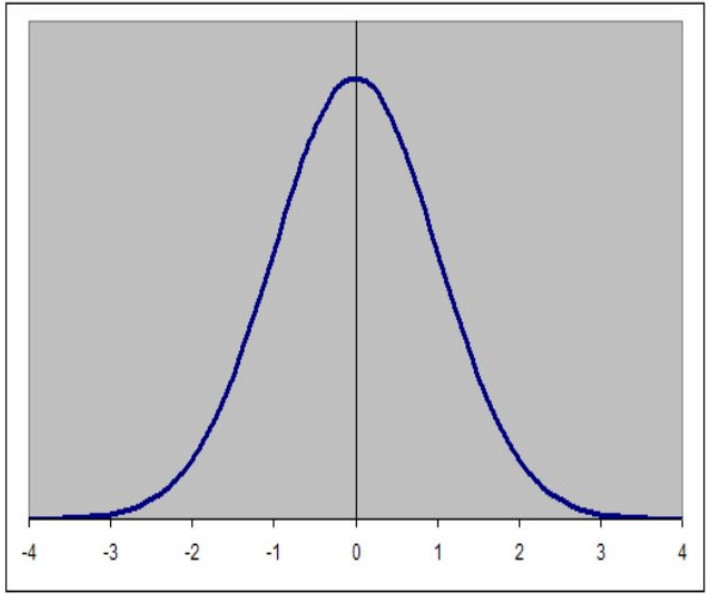
$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$f(x) = \frac{d}{dx}F(x)$$

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

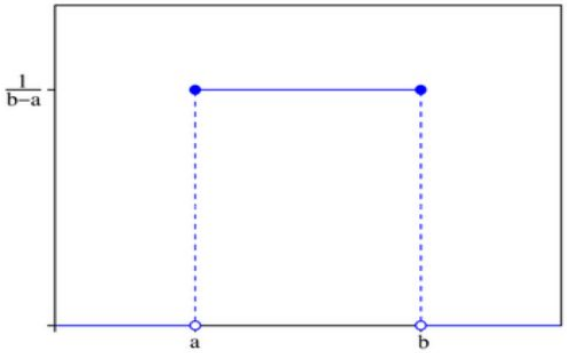
$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx$$

Cumulative Distribution Function

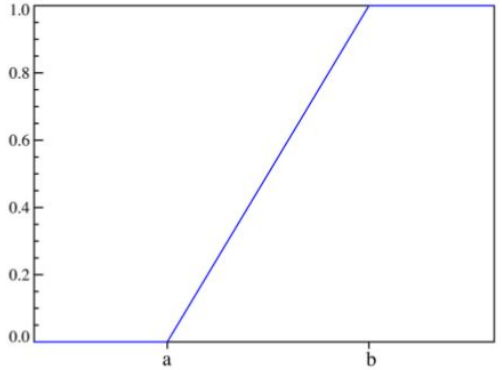


Intuitively, one can think of f(x)dx as being the probability of X falling within the infinitesimal interval [x, x+dx]. $P(x < X < x+dx) = f(x)\,dx$

Uniform Distribution



PDF



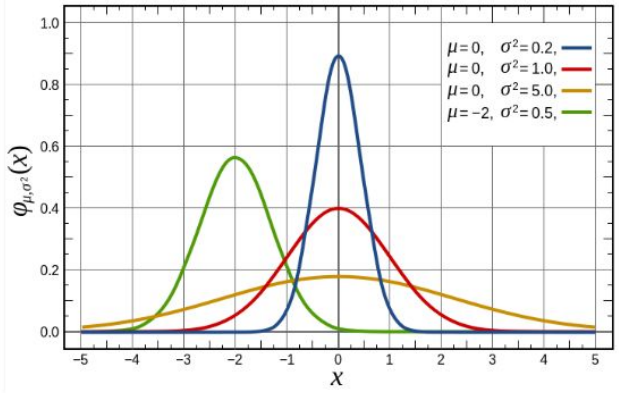
CDF

$$f(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{Otherwise} \end{cases}$$

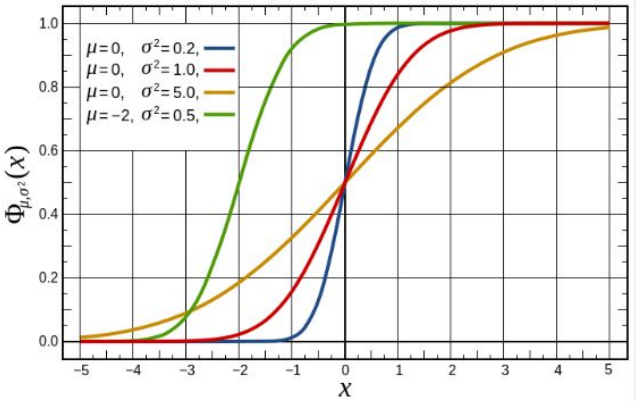
$$F(x) = \begin{cases} 0 & x \le a \\ \frac{x-a}{b-a} & a < x \le b \\ 1 & b < x \end{cases}$$

slide credit: Barnabás Póczos & Alex Smola

Normal (Gaussian) Distribution



PDF



CDF

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

$$F(x) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right)\right]$$

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$$

slide credit: Barnabás Póczos & Alex Smola

Moments

Expectation: average value, mean, 1^{st} moment:

$$\mathbb{E}[X] = \begin{cases} \sum_{i \in \Omega} x_i \, f(x_i) & \text{discrete} \\ \int_{-\infty}^{+\infty} x \, f(x) dx & \text{continuos} \end{cases}$$

f(x) = probability mass

f(x) = probability density

- $\mathbb{E}[a] = a$ for any constant $a \in \mathbb{R}$
- $\mathbb{E}[a\,f(X) + b\,g(X)] = a\mathbb{E}[f(X)] + b\mathbb{E}[g(X)]$ for any constants $a, b \in \mathbb{R}$

Moments

Variance: spread, 2^{st} moment:

$$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \begin{cases} \sum_{i \in \Omega} (x_i - \mathbb{E}[X])^2 f(x_i) & \text{discrete} \\ \int_{-\infty}^{+\infty} (x - \mathbb{E}[X])^2 f(x) dx & \text{continuos} \end{cases}$$

- $\mathrm{Var}[a] = 0$ for any constant $a \in \mathbb{R}$
- $\mathrm{Var}[a\, f(X)] = a^2 \mathrm{Var}[f(X)]$ for any constant $a \in \mathbb{R}$

Slide Credit: Francesco Orabona

MLE for Gaussian mean and variance

Choose $\theta = (\mu, \sigma^2)$ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(x_i|\theta)$$ Independent draws

$$= \arg\max_{\theta} \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$ Identically distributed

$$= \arg\max_{\theta=(\mu,\sigma^2)} \underbrace{(\frac{1}{\sigma\sqrt{2\pi}})^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}}_{J(\theta)}$$

$$\begin{aligned}\log(f(x_1, x_2, \ldots, x_n|\sigma, \mu)) &= \log((\tfrac{1}{\sigma\sqrt{2\pi}})^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}) \\ &= n\log\tfrac{1}{\sigma\sqrt{2\pi}} - \tfrac{1}{2\sigma^2}\textstyle\sum_{i=1}^{n}(x_i-\mu)^2 \\ &= -\tfrac{n}{2}\log(2\pi) - n\log\sigma - \tfrac{1}{2\sigma^2}\textstyle\sum_{i=1}^{n}(x_i-\mu)^2\end{aligned}$$

Let's call $\log(f(x_1, x_2, \ldots, x_n|\sigma, \mu))$ as $\mathcal{L}$, then let:

$$\frac{d\mathcal{L}}{d\mu} = \frac{1}{2\sigma^2}\sum_{i=1}^{n}(2\hat{\mu} - 2x_i) = 0$$

Because σ^2 should be larger than zero,

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

MLE for Gaussian mean and variance

$$\begin{aligned}\log(f(x_1, x_2, \ldots, x_n|\sigma, \mu)) &= \log((\frac{1}{\sigma\sqrt{2\pi}})^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}) \\ &= n\log\frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \\ &= -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\end{aligned}$$

Let's call $\log(f(x_1, x_2, \ldots, x_n|\sigma, \mu))$ as $\mathcal{L}$, then let:

$$\frac{d\mathcal{L}}{d\mu} = \frac{1}{2\sigma^2}\sum_{i=1}^{n}(2\hat{\mu} - 2x_i) = 0$$

Because σ^2 should be larger than zero,

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

MLE for Gaussian mean and variance

Choose $\theta = (\mu, \sigma^2)$ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(x_i|\theta)$$ Independent draws

$$= \arg\max_{\theta} \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$ Identically distributed

$$= \arg\max_{\theta=(\mu,\sigma^2)} \underbrace{\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}}_{J(\theta)}$$

$$\frac{d\mathcal{L}}{d\sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n}(x_i-\mu)^2\sigma^{-3} = 0$$

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

MLE for Gaussian mean and variance

Choose $\theta = (\mu, \sigma^2)$ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(x_i|\theta)$$ Independent draws

$$= \arg\max_{\theta} \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$ Identically distributed

$$= \arg\max_{\theta=(\mu,\sigma^2)} \underbrace{\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}}_{J(\theta)}$$

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

AMML

Exercise on MLE and MAP

Suppose data set D={2,5,9,5,4,8} is an i.i.d. sample from a Poisson distribution with an unknown parameter λ. Find the maximum likelihood estimate of λ

The probability density function of a Poisson distribution is expressed as $p(x|\lambda) = \lambda^x e^{-\lambda}/x!$, with some parameter $\lambda \in \mathbb{R}^+$. We will estimate this parameter as

$$\lambda_{ML} = \underset{\lambda \in (0,\infty)}{\arg\max} \{p(\mathcal{D}|\lambda)\}.$$

Likelihood

$$\begin{aligned} p(\mathcal{D}|\lambda) &= p(\{x_i\}_{i=1}^n |\lambda) \\ &= \prod_{i=1}^n p(x_i|\lambda) \\ &= \frac{\lambda^{\sum_{i=1}^n x_i} \cdot e^{-n\lambda}}{\prod_{i=1}^n x_i!}. \end{aligned}$$

To find λ that maximizes the likelihood, we will

- take the logarithm (a monotonic function) to simplify the calculation,
- find its derivative with respect to λ.
- equate it with zero to find the maximum.

Exercise on MLE and MAP

Suppose data set D={2,5,9,5,4,8} is an i.i.d. sample from a Poisson distribution with an unknown parameter λ. Find the maximum likelihood estimate of λ

The probability density function of a Poisson distribution is expressed as $p(x|\lambda) = \lambda^x e^{-\lambda}/x!$, with some parameter $\lambda \in \mathbb{R}^+$. We will estimate this parameter as

$$\lambda_{ML} = \underset{\lambda \in (0,\infty)}{\arg\max} \{p(\mathcal{D}|\lambda)\}.$$

Likelihood

$$\begin{aligned} p(\mathcal{D}|\lambda) &= p(\{x_i\}_{i=1}^n|\lambda) \\ &= \prod_{i=1}^n p(x_i|\lambda) \\ &= \frac{\lambda^{\sum_{i=1}^n x_i} \cdot e^{-n\lambda}}{\prod_{i=1}^n x_i!}. \end{aligned}$$

Log-likelihood

$$\begin{aligned} ll(D,\lambda) &= \ln p(\mathcal{D}|\lambda) \\ &= \sum_{i=1}^n x_i \ln \lambda - n\lambda - \sum_{i=1}^n \ln(x_i!) \end{aligned}$$

$$\frac{\partial ll(\mathcal{D},\lambda)}{\partial \lambda} = \frac{1}{\lambda}\sum_{i=1}^n x_i - n = 0. \Longrightarrow \lambda_{ML} = \frac{1}{n}\sum_{i=1}^n x_i$$

Exercise on MLE and MAP

Suppose data set D={2,5,9,5,4,8} is an i.i.d. sample from a Poisson distribution with an unknown parameter λ. Find the maximum likelihood estimate of λ

The probability density function of a Poisson distribution is expressed as $p(x|\lambda) = \lambda^x e^{-\lambda}/x!$, with some parameter $\lambda \in \mathbb{R}^+$. We will estimate this parameter as

$$\lambda_{ML} = \underset{\lambda \in (0,\infty)}{\arg\max} \{p(\mathcal{D}|\lambda)\}.$$

By substituting $n = 6$ and values from $\mathcal{D}$, we can compute the solution as

$$\begin{aligned} \lambda_{ML} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ &= 5.5, \end{aligned}$$

Exercise on MLE and MAP

Suppose data set D={2,5,9,5,4,8} is an i.i.d. sample from a Poisson distribution with an unknown parameter λ. But now we are also given additional information: suppose the prior knowledge about λ can be expressed using a gamma distribution $\Gamma(x|k,\theta)$ with parameters k=3, θ=1.
Find the maximum a posteriori estimate of λ

The probability density function of a Poisson distribution is expressed as $p(x|\lambda) = \lambda^x e^{-\lambda}/x!$, with some parameter $\lambda \in \mathbb{R}^+$. We will estimate this parameter as

$$\lambda_{MAP} = \arg\max_{\lambda \in (0,\infty)} \{p(\mathcal{D}|\lambda)p(\lambda)\}$$

Exercise on MLE and MAP

Suppose data set D={2,5,9,5,4,8} is an i.i.d. sample from a Poisson distribution with an unknown parameter λ. But now we are also given additional information: suppose the prior knowledge about λ can be expressed using a gamma distribution $\Gamma(x|k,\theta)$ with parameters k=3, θ=1.
Find the maximum a posteriori estimate of λ

The probability density function of a Poisson distribution is expressed as $p(x|\lambda) = \lambda^x e^{-\lambda}/x!$, with some parameter $\lambda \in \mathbb{R}^+$. We will estimate this parameter as

$$\lambda_{MAP} = \arg\max_{\lambda \in (0,\infty)} \{p(\mathcal{D}|\lambda)p(\lambda)\}$$

First, we write the probability density function of the gamma family as

$$\Gamma(x|k,\theta) = \frac{x^{k-1}e^{-\frac{x}{\theta}}}{\theta^k \Gamma(k)},$$

where $x > 0$, $k > 0$, and $\theta > 0$. $\Gamma(k)$ is the gamma function that generalizes the factorial function; when k is an integer, we have $\Gamma(k) = (k-1)!$.

Exercise on MLE and MAP

Suppose data set D={2,5,9,5,4,8} is an i.i.d. sample from a Poisson distribution with an unknown parameter λ. But now we are also given additional information: suppose the prior knowledge about λ can be expressed using a gamma distribution $\Gamma(x|k,\theta)$ with parameters k=3, θ=1.
Find the maximum a posteriori estimate of λ

The probability density function of a Poisson distribution is expressed as $p(x|\lambda) = \lambda^x e^{-\lambda}/x!$, with some parameter $\lambda \in \mathbb{R}^+$. We will estimate this parameter as

$$\lambda_{MAP} = \arg\max_{\lambda \in (0,\infty)} \{p(\mathcal{D}|\lambda)p(\lambda)\}$$

Likelihood $$p(\mathcal{D}|\lambda) = \frac{\lambda^{\sum_{i=1}^n x_i} \cdot e^{-n\lambda}}{\prod_{i=1}^n x_i!}$$

Prior $$p(\lambda) = \frac{\lambda^{k-1} e^{-\frac{\lambda}{\theta}}}{\theta^k \Gamma(k)}$$

$$\ln p(\lambda|\mathcal{D}) \propto \ln p(\mathcal{D}|\lambda) + \ln p(\lambda)$$
$$= (k - 1 + \sum_{i=1}^n x_i) \ln \lambda - \lambda(n + \frac{1}{\theta}) - \sum_{i=1}^n \ln x_i! - k \ln \theta - \ln \Gamma(k)$$

Exercise on MLE and MAP

Suppose data set D={2,5,9,5,4,8} is an i.i.d. sample from a Poisson distribution with an unknown parameter λ. But now we are also given additional information: suppose the prior knowledge about λ can be expressed using a gamma distribution $\Gamma(x|k,\theta)$ with parameters k=3, θ=1.
Find the maximum a posteriori estimate of λ

The probability density function of a Poisson distribution is expressed as $p(x|\lambda) = \lambda^x e^{-\lambda}/x!$, with some parameter $\lambda \in \mathbb{R}^+$. We will estimate this parameter as

$$\lambda_{MAP} = \arg\max_{\lambda \in (0,\infty)} \{p(\mathcal{D}|\lambda)p(\lambda)\}$$

$$\lambda_{MAP} = \frac{k - 1 + \sum_{i=1}^{n} x_i}{n + \frac{1}{\theta}}$$

$$= 5$$

$$\ln p(\lambda|\mathcal{D}) \propto \ln p(\mathcal{D}|\lambda) + \ln p(\lambda)$$

$$= (k - 1 + \sum_{i=1}^{n} x_i) \ln \lambda - \lambda(n + \frac{1}{\theta}) - \sum_{i=1}^{n} \ln x_i! - k \ln \theta - \ln \Gamma(k)$$

Exercise on MLE and MAP

Suppose data set D={2,5,9,5,4,8} is an i.i.d. sample from a Poisson distribution with an unknown parameter λ.

$$\lambda_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$= 5.5,$$

$$\lambda_{MAP} = \frac{k - 1 + \sum_{i=1}^{n} x_i}{n + \frac{1}{\theta}}$$
$$= 5$$

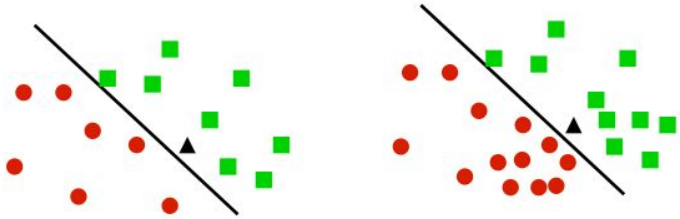
It can be shown that when $n \to \infty$ the expected value of the difference $(\lambda_{MAP}-\lambda_{ML})$ goes to 0.
In other words, large data diminish the importance of prior knowledge.

Discriminative vs Generative

Many supervised learning can be viewed as estimating $P(X,Y)$. Generally thy fall into two categories

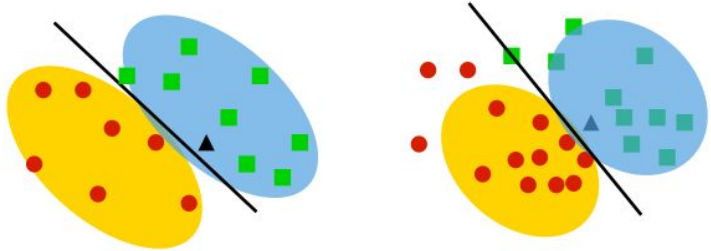
- When we estimate $P(X,Y)=P(X|Y)P(Y)$ then we call it *generative learning.*
- When we only estimate $P(Y|X)$ then we call it *discriminative learning.*

Discriminative



- Only care about estimating the conditional probabilities
- Very good when underlying distribution of data is really complicated (e.g. texts, images, movies)

Generative



- Model observations (x,y) first
- Then infer p(y|x)
- Good for missing variables, better diagnostics
- Easy to add prior knowledge about data

Slide Credit: Barnabás Póczos & Alex Smola

Frequentists vs Bayesian -- MLE vs MAP

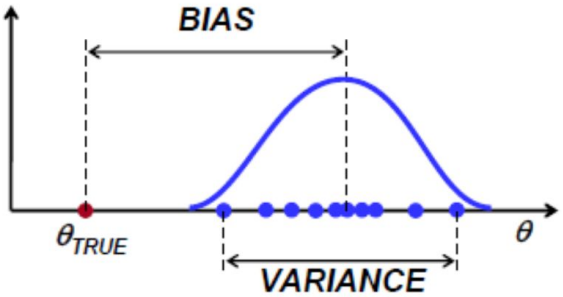
Slide Credit Idea: Barnabás Póczos & Alex Smola

How good is the estimator?

Consider two properties of the estimator:

- Bias
- Variance

  - **BIAS**: how close is the estimate to the true value?
  - **VARIANCE**: how much does the estimate change for different runs (e.g. different datasets)?



The Bias

- The bias of an estimator $\hat{\theta}_m = g(\boldsymbol{x}^{(1)}, \ldots \boldsymbol{x}^{(m)})$ for parameter θ is defined as

$$\text{bias}\left(\hat{\theta}_m\right) = E\left[\hat{\theta}_m\right] - \theta$$

- The estimator is unbiased if $\text{bias}(\hat{\theta}_m) = 0$

Estimator of Bernoulli Mean

- Bernoulli distribution for binary variable $x \in \{0,1\}$ with mean θ has the for $P(x; \theta) = \theta^x (1- \theta)^{1-x}$

- Estimator for given samples $\{x^{(1)}, \ldots, x^{(m)}\}$ is

$$\hat{\theta}_m = \frac{1}{m} \sum_{i=1}^{m} x^{(i)}$$

$$\begin{aligned}
\text{bias}(\hat{\theta}_m) &= E\left[\hat{\theta}_m\right] - \theta \\
&= E\left[\frac{1}{m} \sum_{i-1}^{m} x^{(i)}\right] - \theta \\
&= \frac{1}{m} \sum_{i=1}^{m} E\left[x^{(i)}\right] - \theta \\
&= \frac{1}{m} \sum_{i=1}^{m} \sum_{x^{(i)}=0}^{1} \left(x^{(i)} \theta^{x^{(i)}} (1-\theta)^{(1-x^{(i)})}\right) - \theta \\
&= \frac{1}{m} \sum_{i=1}^{m} (\theta) - \theta = \theta - \theta = 0
\end{aligned}$$

unbiased

The Bias

- The bias of an estimator $\hat{\theta}_m = g(\boldsymbol{x}^{(1)}, \ldots \boldsymbol{x}^{(m)})$ for parameter θ is defined as

$$\text{bias}\left(\hat{\theta}_m\right) = E\left[\hat{\theta}_m\right] - \theta$$

- The estimator is unbiased if $\text{bias}(\hat{\theta}_m) = 0$

Estimator of Gaussian Mean

- $\{x^{(1)}, \ldots, x^{(m)}\}$ distributed i.i.d. according to $p(x^i) = N(x^i; \mu, \sigma^2)$
- sample mean is an estimator of the mean parameter

$$\hat{\mu}_m = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$$

$$\begin{aligned}
\text{bias}(\hat{\mu}_m) &= \mathbb{E}[\hat{\mu}_m] - \mu \\
&= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right] - \mu \\
&= \left(\frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\left[x^{(i)}\right]\right) - \mu \\
&= \left(\frac{1}{m}\sum_{i=1}^{m} \mu\right) - \mu \\
&= \mu - \mu = 0 \qquad \textbf{unbiased}
\end{aligned}$$

The Bias

- The bias of an estimator $\hat{\theta}_m = g(\boldsymbol{x}^{(1)}, \ldots \boldsymbol{x}^{(m)})$ for parameter θ is defined as

$$\text{bias}\left(\hat{\theta}_m\right) = E\left[\hat{\theta}_m\right] - \theta$$

- The estimator is unbiased if $\text{bias}(\hat{\theta}_m) = 0$

Estimator of Gaussian Variance

- The sample variance is $\hat{\sigma}_m^2 = \frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m\right)^2$
- We are interested in computing $\text{bias}(\hat{\sigma}_m^2) = \mathrm{E}(\hat{\sigma}_m^2) - \sigma^2$
- We begin by evaluating

$$\mathbb{E}[\hat{\sigma}_m^2] = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m\right)^2\right]$$

$$= \frac{m-1}{m}\sigma^2$$

The Bias

- The bias of an estimator $\hat{\theta}_m = g(\boldsymbol{x}^{(1)}, \ldots \boldsymbol{x}^{(m)})$ for parameter θ is defined as

$$\text{bias}\left(\hat{\theta}_m\right) = E\left[\hat{\theta}_m\right] - \theta$$

- The estimator is unbiased if $\text{bias}(\hat{\theta}_m) = 0$

Estimator of Gaussian Variance

- The sample variance is $\hat{\sigma}_m^2 = \frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m\right)^2$ **biased**
- We are interested in computing $\text{bias}(\hat{\sigma}_m^2) = \mathrm{E}(\hat{\sigma}_m^2) - \sigma^2$
- We begin by evaluating

$$\mathbb{E}[\hat{\sigma}_m^2] = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m\right)^2\right]$$

$$= \frac{m-1}{m}\sigma^2$$

The unbiased sample variance estimator is

$$\hat{\sigma}_m^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m\right)^2$$

The Variance

- The variance indicate how much we expect the estimator to vary as a function of data samples
- Just as we computed the expectation of the estimator to determine its bias, we can compute its variance
- The variance of an estimator is simply Var($\hat{\theta}$) where the random variable is the training set.

  If two estimators of a parameters are both unbiased, the best is the one with the least amount of variability.

The estimator $\hat{\theta}_1$ is said to be **more efficient** than $\hat{\theta}_2$ if

$$\mathrm{Var}(\hat{\theta}_1) < \mathrm{Var}(\hat{\theta}_2).$$

The Variance

- The variance indicate how much we expect the estimator to vary as a function of data samples
- Just as we computed the expectation of the estimator to determine its bias, we can compute its variance
- The variance of an estimator is simply Var($\hat{\theta}$) where the random variable is the training set.
- The **square root of the variance is called the standard error**, denoted as SE($\hat{\theta}$)
- SE($\hat{\theta}$) measures how we would expect the estimate to vary as we obtained different samples from the same distribution
- The **standard error of the mean** is given by

$$SE\left(\hat{\mu}_m\right) = \sqrt{Var\left[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right]} = \frac{\sigma}{\sqrt{m}}$$

Here σ is the true variance of the samples x(i).
The SE is often estimated using an estimate of σ: it is not unbiased, but the approximation is reasonable

Mean Squared Error

Definition: Mean Squared Error

Let $\widehat{\theta}$ be an estimator for an unknown parameter θ. The **Mean Squared Error (MSE)** is defined as

$$\mathrm{MSE}(\widehat{\theta}) = \mathbb{E}[(\widehat{\theta} - \theta)^2] \ .$$

Clearly, if this is zero then the estimator is perfect. There are of course other ways of measuring the error of an estimator, e.g.

$$\mathrm{MAE}(\widehat{\theta}) = \mathbb{E}_\theta[|\widehat{\theta} - \theta|] \quad \text{(Mean Absolute Error)}$$

$$\mathbb{P}(|\widehat{\theta} - \theta| \geq \epsilon), \ \text{for } \epsilon > 0$$

Slide Credit: Rui Castro

Bias / Variance Decomposition

Theorem: Bias-Variance Decomposition

Let $\hat{\theta}$ be an estimator for an unknown parameter θ. The mean squared error of this estimator can be written in terms of the bias and variance as follows:

$$\text{MSE}(\hat{\theta}) = b^2(\hat{\theta}) + \text{Var}(\hat{\theta}) \ .$$

Proof:

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] \\
&= \mathbb{E}\left[(\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta)^2\right] \\
&= \mathbb{E}\left[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2 + (\mathbb{E}[\hat{\theta}] - \theta)^2 + 2(\hat{\theta} - \mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}] - \theta)\right] \\
&\qquad \text{not random!} \\
&= \mathbb{E}\left[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2\right] + (\mathbb{E}[\hat{\theta}] - \theta)^2 + 2(\mathbb{E}[\hat{\theta}] - \mathbb{E}[\hat{\theta}])\left(\mathbb{E}[\hat{\theta}] - \theta\right) \\
&\qquad = 0 \\
&= \text{Var}(\hat{\theta}) + b^2(\hat{\theta})
\end{aligned}$$

22

Example - Rate of a Poisson Process

It is reasonable to assume that users access a webserver according to a Poisson process with unknown rate λ. Suppose you measure the number of users in 4 one hour periods:

$$X_1, \ldots, X_4 \qquad \text{where } X_i \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$$

Two estimators are proposed, which one is better?

$$\hat{\lambda}_1 = \frac{X_1 + 2X_2 + 3X_3 + 4X_4}{10} \qquad \hat{\lambda}_2 = \frac{1 + X_1 + X_2 + X_3 + X_4}{4}$$

Slide Credit: Rui Castro

Example - Rate of a Poisson Process

Let's first see what the bias of the estimators is

$$\begin{aligned}\mathbb{E}[\hat{\lambda}_1] &= \mathbb{E}\left[\frac{X_1 + 2X_2 + 3X_3 + 4X_4}{10}\right] \\ &= \frac{1}{10}\mathbb{E}[X_1] + \frac{2}{10}\mathbb{E}[X_2] + \frac{3}{10}\mathbb{E}[X_3] + \frac{4}{10}\mathbb{E}[X_4] = \lambda\end{aligned}$$

So $\text{bias}(\hat{\lambda}_1) = \mathbb{E}[\hat{\lambda}_1] - \lambda = 0$ (unbiased estimator)

$$\begin{aligned}\text{bias}(\hat{\lambda}_2) = \mathbb{E}[\hat{\lambda}_2] - \lambda &= \mathbb{E}\left[\frac{1 + X_1 + X_2 + X_3 + X_4}{4}\right] - \lambda \\ &= ???\end{aligned}$$

A - 0
B - $\frac{4}{5}\lambda$
C - $-1/4$
D - $1/4$

Slide Credit: Rui Castro

Example - Rate of a Poisson Process

Let's first see what the bias of the estimators is

$$\begin{aligned}\mathbb{E}[\hat{\lambda}_1] &= \mathbb{E}\left[\frac{X_1 + 2X_2 + 3X_3 + 4X_4}{10}\right] \\ &= \frac{1}{10}\mathbb{E}[X_1] + \frac{2}{10}\mathbb{E}[X_2] + \frac{3}{10}\mathbb{E}[X_3] + \frac{4}{10}\mathbb{E}[X_4] = \lambda\end{aligned}$$

$$\text{So bias}(\hat{\lambda}_1) = \mathbb{E}[\hat{\lambda}_1] - \lambda = 0 \text{ (unbiased estimator)}$$

$$\text{bias}(\hat{\lambda}_2) = \mathbb{E}[\hat{\lambda}_2] - \lambda = \mathbb{E}\left[\frac{1 + X_1 + X_2 + X_3 + X_4}{4}\right] - \lambda$$

$$= \frac{1}{4} + \frac{4\lambda}{4} - \lambda = \frac{1}{4}$$

It seems the first estimator might be more desirable...

Slide Credit: Rui Castro

Example - Rate of a Poisson Process

$$V[\hat{\lambda}_1] = V\left(\frac{X_1 + 2X_2 + 3X_3 + 4X_4}{10}\right)$$
$$= \frac{1}{100}V(X_1) + \frac{4}{100}V(X_2) + \frac{9}{100}V(X_3) + \frac{16}{100}V(X_4) = \frac{3}{10}\lambda$$

So $\text{MSE}(\hat{\lambda}_1) = \text{bias}^2(\hat{\lambda}_1) + V(\hat{\lambda}_1) = \frac{3}{10}\lambda$

$$V[\hat{\lambda}_2] = V\left(\frac{1 + X_1 + X_2 + X_3 + X_4}{4}\right) = ???$$

A - $\frac{1}{4} + \lambda$
B - $\frac{1}{4}(1 + \lambda)$
C - $\lambda/4$
D - $1 + \lambda/4$

Slide Credit: Rui Castro

Example - Rate of a Poisson Process

$$V[\hat{\lambda}_1] = V\left(\frac{X_1 + 2X_2 + 3X_3 + 4X_4}{10}\right)$$
$$= \frac{1}{100}V(X_1) + \frac{4}{100}V(X_2) + \frac{9}{100}V(X_3) + \frac{16}{100}V(X_4) = \frac{3}{10}\lambda$$

So $\text{MSE}(\hat{\lambda}_1) = \text{bias}^2(\hat{\lambda}_1) + V(\hat{\lambda}_1) = \frac{3}{10}\lambda$

$$V[\hat{\lambda}_2] = V\left(\frac{1 + X_1 + X_2 + X_3 + X_4}{4}\right) = ???$$
$$= \frac{\lambda}{4}$$

So $\text{MSE}(\hat{\lambda}_2) = \text{bias}^2(\hat{\lambda}_2) + V(\hat{\lambda}_2) = \frac{1}{16} + \frac{1}{4}\lambda$

Slide Credit: Rui Castro

Example - Rate of a Poisson Process

$$\text{So } \text{MSE}(\hat{\lambda}_1) = \text{bias}^2(\hat{\lambda}_1) + V(\hat{\lambda}_1) = \frac{3}{10}\lambda$$

$$\text{So } \text{MSE}(\hat{\lambda}_2) = \text{bias}^2(\hat{\lambda}_2) + V(\hat{\lambda}_2) = \frac{1}{16} + \frac{1}{4}\lambda$$

Which estimator should we prefer?

If $\lambda \geq \frac{5}{4} = 1.25$ then $\text{MSE}(\hat{\lambda}_1) \geq \text{MSE}(\hat{\lambda}_2)$ so the second estimator is better. Otherwise the first estimator is better.

The answer depends on the unknown value of lambda. If someone tells us that every hour hundreds of people access the webserver then we know which estimator to choose.

Slide Credit: Rui Castro

Summary

- Maximum a Posteriori
- Maximum Likelihood
- iid data
- Discrete and Continuous Distributions
- Bias / Variance
- Mean Squared Error

Expected Values and Variance

Distribution	PDF or PMF	Mean	Variance
Bernoulli(p)	$\begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0. \end{cases}$	p	$p(1-p)$
Binomial(n, p)	$\binom{n}{k} p^k (1-p)^{n-k}$ for $k = 0, 1, \ldots, n$	np	$np(1-p)$
Geometric(p)	$p(1-p)^{k-1}$ for $k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson(λ)	$\frac{e^{-\lambda}\lambda^k}{k!}$ for $k = 0, 1, \ldots$	λ	λ
Uniform(a, b)	$\frac{1}{b-a}$ for all $x \in (a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Gaussian(μ, σ^2)	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ for all $x \in (-\infty, \infty)$	μ	σ^2
Exponential(λ)	$\lambda e^{-\lambda x}$ for all $x \geq 0, \lambda \geq 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Slide Credit: cs229-prob stanford