Artificial Intelligence and Machine Learning

Barbara Caputo

Useful info (1)

Teacher: Barbara Caputo,
https://scholar.google.it/citations
?hl=en&user=mHbdIAwAAAAJ&view_op=list_works

Assistant: Paolo Russo

Where/how to find us:

Email: barbara.caputo@iit.it
paolo.russo@iit.it

Office: 4E-45 (BC), Lab 7 –R3BC31 (PR)

Q/A time: after the lectures (BC)/by appointment
You can come any time at your own risk!

Useful info (1)

Textbooks/Useful books:

K. P. Murphy. Machine Learning: a probabilistic perspective. MIT Press.

B. Schoelkopf, A. Smola. Learning with Kernels. MIT Press.

W. Feller. An introduction to probability theory and its applications. Vol 1. Wiley Ed.

I. Goodfellows, Y. Bengio, A. Courville. Deep Learning. MIT Press.

Useful info (2)

Exam modality:

Homeworks (lab experiences) +written exam +oral exam

Useful info (2)

Exam modality:

-Homeworks during the lectures: 3, on three key topics (PCA, svm, cnn???e-m???). Reports to be submitted by end of the lectures (19/01/2019, 23:59 CET). Every experience graded (2 points), if overall points <4 no admission to written exam.

Useful info (2)

Exam modality:

-Homeworks during the lectures: 3, on three key topics (PCA, svm, cnn???e-m???). Reports to be submitted by end of the lectures (19/01/2019, 23:59 CET). Every experience graded (2 points), if overall points <4 no admission to written exam.

- there will be a cut-off date for submitting the reports for every exam session

Useful info (2)

Exam modality:

-Homeworks during the lectures: 3, on three key topics (PCA, svm, cnn???e-m???). Reports to be submitted by end of the lectures (19/01/2019, 23:59 CET). Every experience graded (2 points), if overall points <4 no admission to written exam.

- Written exam: exercises as done during lectures. If grade <18, no admission to oral exam.

Useful info (2)

Exam modality:

-Homeworks during the lectures: 3, on three key topics (PCA, svm, cnn???). Reports to be submitted by end of the lectures (19/01/2019, 23:59 CET). Every experience graded (2 points), if overall points <4 no admission to written exam.

- Written exam: exercises as done during lectures. If grade <18, no admission to oral exam.

- Oral exam: pen and pencil, three questions. You decide the first, the second is decided the day before the exam, the third I decide.

Useful info (2)

Exam modality:

-Homeworks during the lectures: 3, on three key topics (PCA, svn, cnn???). Reports to be submitted by end of the lectures (19/01/2019, 23:59 CET). Every experience graded (2 points), if overall points <4 no admission to written exam.

- Written exam: exercises as done during lectures. If grade <18, no admission to oral exam.

- Oral exam: pen and pencil, three questions. You decide the first, the second is decided the day before the exam, the third I decide.

Useful info (2)

Exam modality:

- Written exam: exercises as done during lectures. If grade <18, no admission to oral exam.

- Oral exam: pen and pencil, three questions. You decide the first, the second is decided the day before the exam, the third I decide. Final grade average between written and oral exam.

Artificial Intelligence and Machine Learning

AI: theory and algorithms that enable computers to mimic human intelligence

ML: a subset of AI that includes statistical techniques enabling machines to improve at tasks with experience.

This course focuses on ML only, and specifically on a specific family of ML algorithms (discriminative methods)

Outline

Theory:

- Probabilities:
  - Probability measures, events, random variables, conditional probabilities, dependence, expectations, etc
- Bayes rule
- Parameter estimation:
  - Maximum Likelihood Estimation (MLE)
  - Maximum a Posteriori (MAP)

Sample space

Def: 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*:*

–Ω may be the set of all possible outcomes of a dice roll (1,2,3,4,5,6)

-Pages of a book opened randomly. (1-157)

-Real numbers for temperature, location, time, etc

Events

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

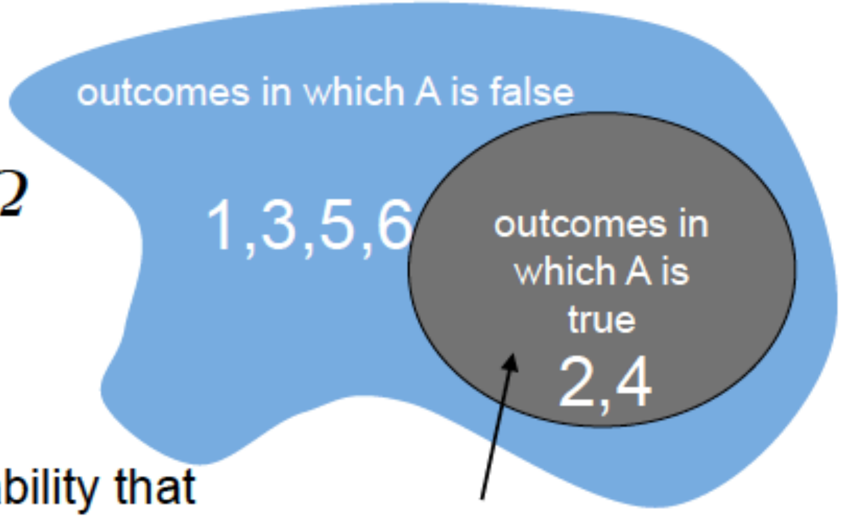
Def: *Event* *A is a* ***subset*** *of the sample space* Ω

Examples:
What is the probability of
- *the book is open at an odd number*
- *rolling a dice the number* <4
- *a random person's height* X *:* $a<X<b$

Probability

Def: *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*.

sample space Ω



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

P(A) is the volume of the area.

Kolmogorov Axioms

(i) Nonnegativity: $P(A) \geq 0$ for each A event.

(ii) $P(\Omega) = 1$.

(iii) σ-additivity: For disjoint sets (events) A_i, we have

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

Consequences:

$P(\emptyset) = 0$.

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

$P(A^c) = 1 - P(A)$.

Random Variables

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

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

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

Examples:

- **Discrete random variable examples (Ω is discrete):**
- $X(\omega)$ = True if a randomly drawn person (ω) from our class (Ω) is female
- $X(\omega)$ = The hometown $X(\omega)$ of a randomly drawn person (ω) from our class (Ω)

Discrete Distributions

- Bernoulli distribution: Ber(p)

$\Omega = \{\text{head, tail}\}$ $X(head) = 1$, $X(tail) = 0$.

$$P(X=a) = P(\omega : X(\omega) = a) = \begin{cases} p, & \text{for } a = 1 \\ 1-p, & \text{for } a = 0 \end{cases}$$

- Binomial distribution: Bin(n,p)

Suppose a coin with head prob. *p* is tossed *n* times. What is the probability of getting *k* heads and *n-k* tails?

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

$K(\omega) = \text{number of heads in } \omega = (\omega_1, \ldots, \omega_n) \in \{\text{head, tail}\}^n = \Omega$

$$P(K=k) = P(\omega : K(\omega) = k) = \sum_{\omega : K(\omega) = k} p^k (1-p)^{n-k} = \binom{n}{k} p^k (1-p)^{n-k}$$

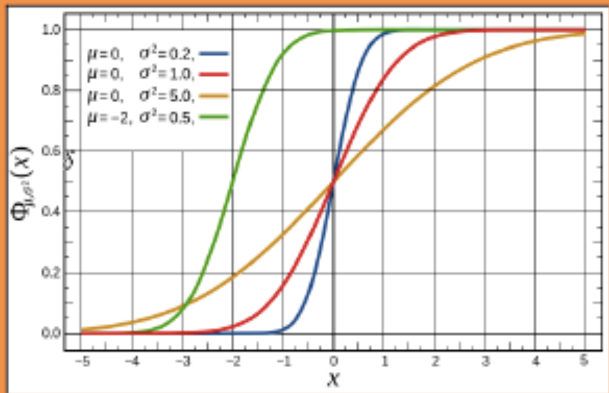
Continuous Distribution

Def: continuous probability distribution: its **cumulative distribution function** is **absolutely continuous**.

Def: cumulative distribution function

USA: $F_X(z) = P(X \leq z)$

Hungary: $F_X(z) = P(X < z)$



Def : Let $F(-\infty) = 0$. $F : (-\infty, \infty) \to \mathbb{R}$ is **absolutely continuous**

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

Def : f is called the density of the distribution.

Properties : $\frac{d}{dx}F(x) = f(x)$ $F(x) = \int_{-\infty}^{x} f(t)dt$

Probability Density Function (pdf)

Pdf properties:

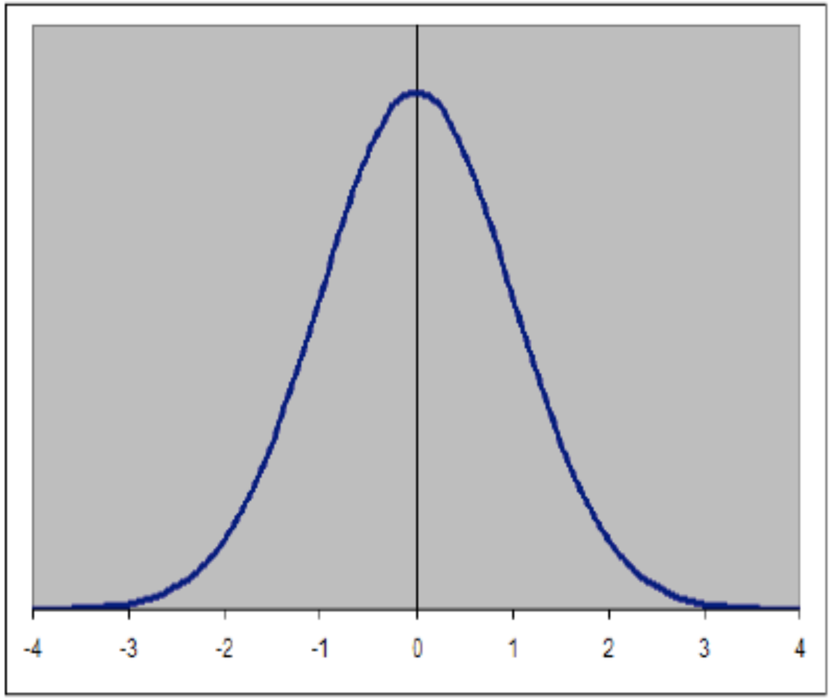
$$f(x) \geq 0$$

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

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

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

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



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

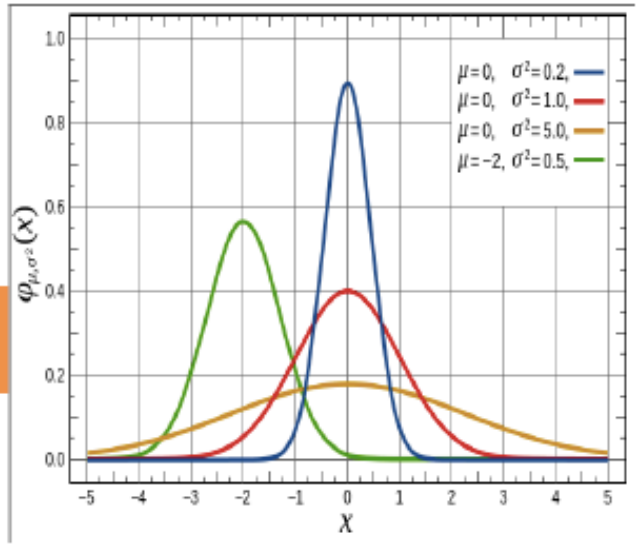
Moments

Expectation: average value, mean, 1st moment:

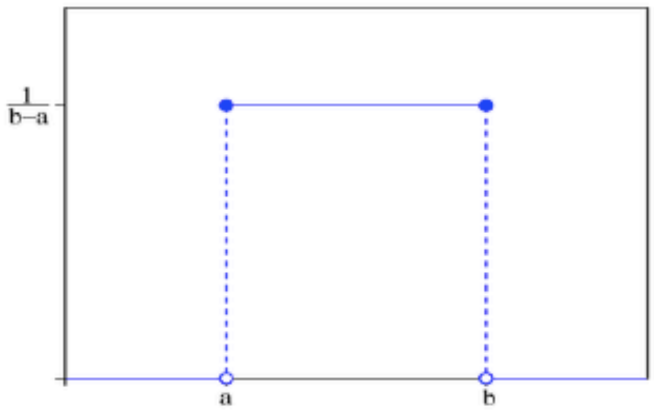
$$E(X) = \begin{cases} \sum_{i \in \Omega} x_i p(x_i) & \text{discrete} \\ \int_{-\infty}^{\infty} x p(x) dx & \text{continuous} \end{cases}$$

Variance: the spread, 2nd moment:

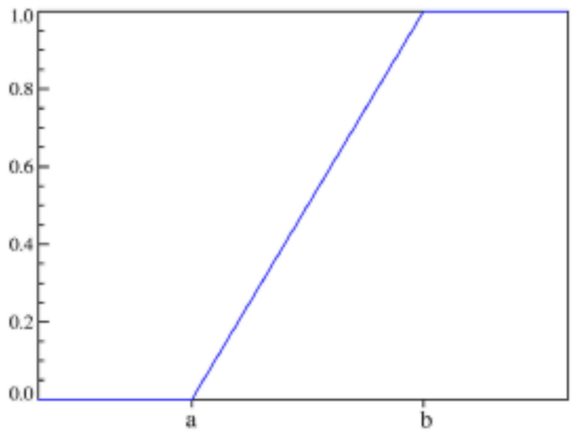
$$E(X) = \begin{cases} \sum_{i \in \Omega} [x_i - E(X)]^2 p(x_i) & \text{discrete} \\ \int_{-\infty}^{\infty} (x - E(x))^2 p(x) dx & \text{continuous} \end{cases}$$



Uniform Distribution



PDF

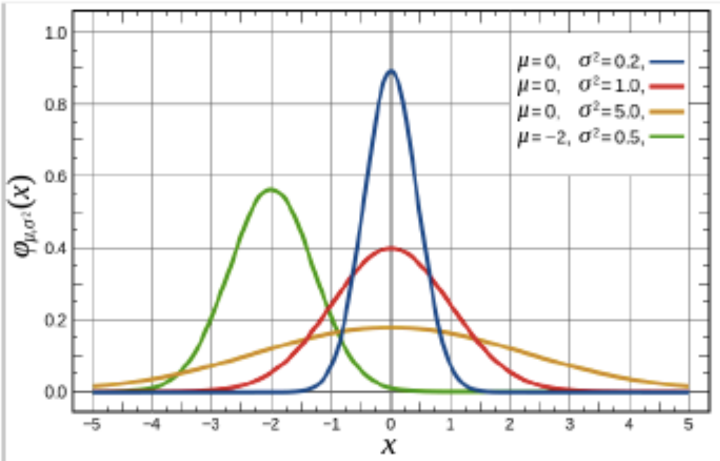


CDF

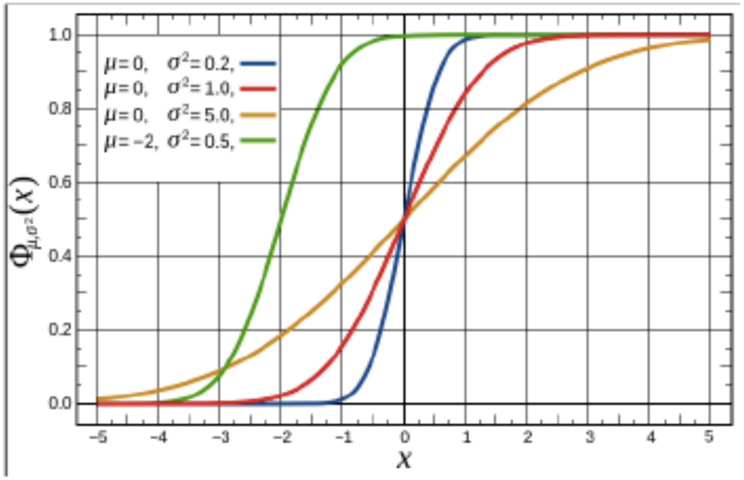
$$f(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{Otherwise} \end{cases}$$

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

Normal (Gaussian) Distribution



PDF



CDF

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

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

Multivariate (Joint) Distribution

Multivariate (Joint) Distribution

We can generalize the above ideas from 1-dimension to any finite dimensions.

$P(a \le X \le b, c \le Y \le d) = ?$ $\qquad$ $P(a_1 \le X_1 \le b_1, \ldots a_d \le X_d \le b_d) = ?$

Multivariate (Joint) Distribution

We can generalize the above ideas from 1-dimension to any finite dimensions.

$P(a \leq X \leq b, c \leq Y \leq d) =?$

$P(a_1 \leq X_1 \leq b_1, \ldots a_d \leq X_d \leq b_d) =?$

Multivariate (Joint) Distribution

We can generalize the above ideas from 1-dimension to any finite dimensions.

$P(a \leq X \leq b, c \leq Y \leq d) = ?$ $\quad$ $P(a_1 \leq X_1 \leq b_1, \ldots a_d \leq X_d \leq b_d) = ?$

Discrete distribution:

$P(\text{headache} \wedge \text{no flu}) = 7/80$
$P(\text{headache}) = 7/80 + 1/80$

$P(X = \text{headache}, Y = \text{flu}) = 1/80$

	Flu	No Flu
Headache	1/80	7/80
No Headache	1/80	71/80

Multivariate Gaussian distribution

Multivariate Gaussian distribution

For $A \subset \mathbb{R}^d$, $P([X_1, \ldots, X_d] \in A) = \int_A f(x_1, \ldots, x_d) dx_1 \cdots dx_d$

$F_X(z_1, \ldots, z_d) = \int_{-\infty}^{z_1} \cdots \int_{-\infty}^{z_d} f(x_1, \ldots, x_d) dx_1 \cdots dx_d$ Multivariate CDF

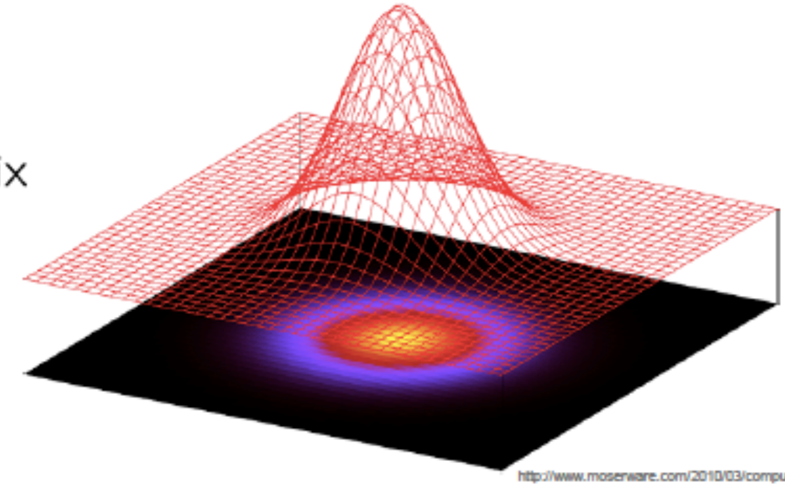
Multivariate Gaussian distribution

For $A \subset \mathbb{R}^d$, $P([X_1, \ldots, X_d] \in A) = \int_A f(x_1, \ldots, x_d) dx_1 \cdots dx_d$

$$F_X(z_1, \ldots, z_d) = \int_{-\infty}^{z_1} \cdots \int_{-\infty}^{z_d} f(x_1, \ldots, x_d) dx_1 \cdots dx_d$$ Multivariate CDF

$\mu \in \mathbb{R}^d$: mean vector

$\Sigma \in \mathbb{R}^{d \times d}$: covariance matrix



http://www.moserware.com/2010/03/computing-your-skill.htm

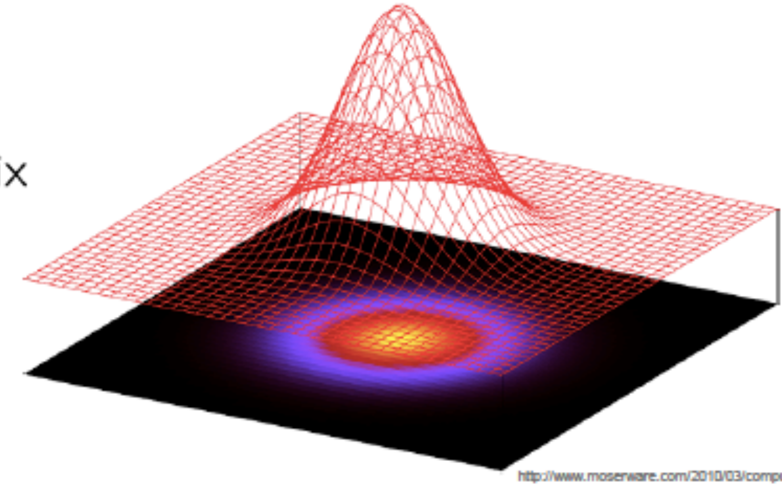
Multivariate Gaussian distribution

For $A \subset \mathbb{R}^d$, $P([X_1, \ldots, X_d] \in A) = \int_A f(x_1, \ldots, x_d) dx_1 \cdots dx_d$

$F_X(z_1, \ldots, z_d) = \int_{-\infty}^{z_1} \cdots \int_{-\infty}^{z_d} f(x_1, \ldots, x_d) dx_1 \cdots dx_d$ Multivariate CDF

$\mu \in \mathbb{R}^d$: mean vector

$\Sigma \in \mathbb{R}^{d \times d}$: covariance matrix



http://www.moserware.com/2010/03/computing-your-skill.htm

$$f_X(x_1, \ldots, x_d) = \frac{1}{\sqrt{|2\pi\Sigma|}} \exp\left(-\frac{1}{2}(\mathrm{x} - \mu)^T \Sigma^{-1} (\mathrm{x} - \mu)\right)$$

Conditional Probability

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

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

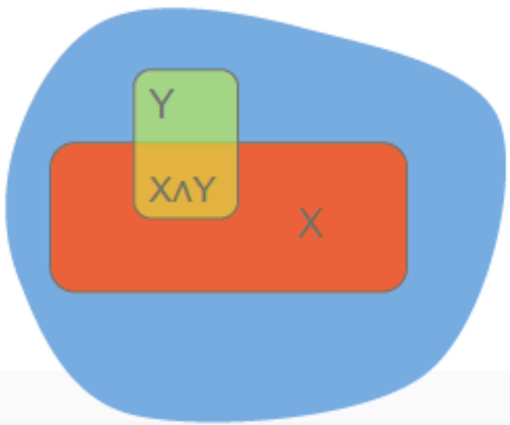
Conditional Probability

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

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

$$P(\text{flu}|\text{headache}) = \frac{P(\text{flu, headache})}{P(\text{headache})} = \frac{1/80}{1/80 + 7/80}$$

	Flu	No Flu
Headache	1/80	7/80
No Headache	1/80	71/80



Independence

Independence

Independent random variables:

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

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

Independence

Independent random variables:

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

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

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

Examples:

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

Conditionally Independent

Conditionally Independent

Conditionally independent:

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

Knowing Z makes X and Y independent

Conditionally Independent

Conditionally independent:

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

Knowing Z makes X and Y independent

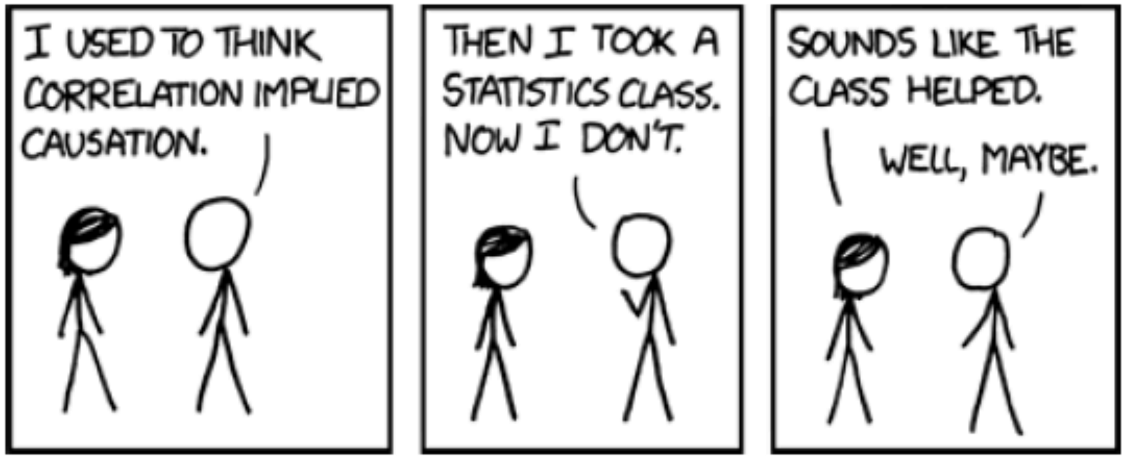
Examples:

Dependent: show size and reading skills
Conditionally independent: show size and reading skills given age

Conditionally 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…



xkcd.com

Conditional Independence

Formally: X is **conditionally independent** of Y given Z:

Conditional Independence

Formally: X is **conditionally independent** of Y given Z:

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

Conditional Independence

Formally: X is **conditionally independent** of Y given Z:

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

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

Conditional Independence

Formally: X is **conditionally independent** of Y given Z:

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

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

Equivalent to:

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

Let’s take a break!

Bayes Rule

network.

Chain Rule & Bayes Rule

Chain Rule & Bayes Rule

Chain rule:

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

Chain Rule & Bayes Rule

Chain rule:

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

Bayes rule:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

Chain Rule & Bayes Rule

Chain rule:

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

Bayes rule:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

Bayes rule is important for reverse conditioning.

AIDS test (Bayes rule)

Data

- Approximately 0.1% are infected
- Test detects all infections
- Test reports positive for 1% healthy people

AIDS test (Bayes rule)

Data

- Approximately 0.1% are infected
- Test detects all infections
- Test reports positive for 1% healthy people

Probability of having AIDS if test is positive:

$$P(a=1|t=1) = \frac{P(t=1|a=1)P(a=1)}{P(t=1)}$$

$$= \frac{P(t=1|a=1)P(a=1)}{P(t=1|a=1)P(a=1) + P(t=1|a=0)P(a=0)}$$

$$= \frac{1 \cdot 0.001}{1 \cdot 0.001 + 0.01 \cdot 0.999} = 0.091$$

Only 9%!...

Improving the diagnosis

Use a follow-up test!

- Test 2 reports positive for 90% infections
- Test 2 reports positive for 5% healthy people

Improving the diagnosis

Use a follow-up test!

- Test 2 reports positive for 90% infections
- Test 2 reports positive for 5% healthy people

$$P(a=0|t_1=1,t_2=1) = \frac{P(t_1=1,t_2=1|a=0)P(a=0)}{P(t_1=1,t_2=1|a=1)P(a=1)+P(t_1=1,t_2=1|a=0)P(a=0)}$$

$$= \frac{0.01 \cdot 0.05 \cdot 0.999}{1 \cdot 0.9 \cdot 0.001 + 0.01 \cdot 0.05 \cdot 0.999} = 0.357$$

$$P(a=1|t_1=1,t_2=1) = 0.643$$

Why can't we use Test 1 twice?

Outcomes are **not** independent but tests 1 and 2 are **conditionally independent** $p(t_1,t_2|a) = p(t_1|a) \cdot p(t_2|a)$

Naïve Bayes Assumption

Naïve Bayes Assumption

Naïve Bayes assumption: Features X_1 and X_2 are conditionally independent given the class label Y:

$$P(X_1, X_2|Y) = P(X_1|Y)P(X_2|Y)$$

Naïve Bayes Assumption

Naïve Bayes assumption: Features X_1 and X_2 are conditionally independent given the class label Y:

$$P(X_1, X_2|Y) = P(X_1|Y)P(X_2|Y)$$

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

Naïve Bayes Assumption

Naïve Bayes assumption: Features X_1 and X_2 are conditionally independent given the class label Y:

$$P(X_1, X_2|Y) = P(X_1|Y)P(X_2|Y)$$

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

How many parameters to estimate?

(X is composed of d binary features, e.g. presence of "earn"
Y has K possible class labels)

$\boxed{(2^d\text{-}1)K \text{ vs } (2\text{-}1)dK}$

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, we have the conditional likelihood $P(X_i|Y)$

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, we have the conditional likelihood $P(X_i | Y)$

Decision rule:

$$
\begin{aligned}
f_{NB}(\mathbf{x}) &= \arg\max_y \; P(x_1, \ldots, x_d \mid y)P(y) \\
&= \arg\max_y \prod_{i=1}^{d} P(x_i|y)P(y)
\end{aligned}
$$

Naïve Bayes Algorithm for discrete features

Naïve Bayes Algorithm for discrete features

Training Data: $\{(X^{(j)}, Y^{(j)})\}_{j=1}^{n}$ $\quad X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)})$

n d dimensional features + class labels

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

We need to estimate these probabilities!

Naïve Bayes Algorithm for discrete features

Training Data: $\{(X^{(j)}, Y^{(j)})\}_{j=1}^{n}$ $X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)})$

n d dimensional features + class labels

$$f_{NB}(\mathbf{x}) = \arg\max_y \prod_{i=1}^{d} P(x_i|y)P(y)$$ **We need to estimate these probabilities!**

Estimate them with Relative Frequencies!

For Class Prior $\widehat{P}(y) = \frac{\{\#j : Y^{(j)} = y\}}{n}$

For Likelihood $\frac{\widehat{P}(x_i, y)}{\widehat{P}(y)} = \frac{\{\#j : X_i^{(j)} = x_i, Y^{(j)} = y\}/n}{\{\#j : Y^{(j)} = y\}/n}$

Naïve Bayes Algorithm for discrete features

Training Data: $\{(X^{(j)}, Y^{(j)})\}_{j=1}^{n}$ $\quad X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)})$

n d dimensional features + class labels

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

We need to estimate these probabilities!

Estimate them with Relative Frequencies!

For Class Prior $\quad \widehat{P}(y) = \frac{\{\#j : Y^{(j)} = y\}}{n}$

For Likelihood $\quad \frac{\widehat{P}(x_i, y)}{\widehat{P}(y)} = \frac{\{\#j : X_i^{(j)} = x_i, Y^{(j)} = y\}/n}{\{\#j : Y^{(j)} = y\}/n}$

NB Prediction for test data:

$X = (x_1, \ldots, x_d)$

$$Y = \arg\max_y \widehat{P}(y) \prod_{i=1}^{d} \frac{\widehat{P}(x_i, y)}{\widehat{P}(y)}$$

Subtlety: Insufficient training data

What if you never see a training instance where $X_1 = a$ when $Y = b$?

For example,

there is no X_1='Earn' when Y='SpamEmail' in our dataset.

$$\Rightarrow P(X_1 = a, Y = b) = 0 \Rightarrow P(X_1 = a|Y = b) = 0$$

Subtlety: Insufficient training data

What if you never see a training instance where $X_1 = a$ when $Y = b$?

Subtlety: Insufficient training data

What if you never see a training instance where $X_1 = a$ when $Y = b$?

For example,

there is no X_1=‘Earn’ when Y=‘SpamEmail’ in our dataset.

$$\Rightarrow P(X_1 = a, Y = b) = 0 \Rightarrow P(X_1 = a|Y = b) = 0$$

Subtlety: Insufficient training data

What if you never see a training instance where $X_1 = a$ when $Y = b$?

For example,

there is no X_1='Earn' when Y='SpamEmail' in our dataset.

$$\Rightarrow P(X_1 = a, Y = b) = 0 \Rightarrow P(X_1 = a|Y = b) = 0$$

$$\boxed{\Rightarrow P(X_1 = a, X_2 ... X_n|Y) = P(X_1 = a|Y) \prod_{i=2}^{d} P(X_i|Y) = 0}$$

Subtlety: Insufficient training data

What if you never see a training instance where $X_1 = a$ when $Y = b$?

For example,

there is no X_1='Earn' when Y='SpamEmail' in our dataset.

$$\Rightarrow P(X_1 = a, Y = b) = 0 \Rightarrow P(X_1 = a|Y = b) = 0$$

$$\boxed{\Rightarrow P(X_1 = a, X_2 ... X_n|Y) = P(X_1 = a|Y) \prod_{i=2}^{d} P(X_i|Y) = 0}$$

Thus, no matter what the values $X_2, \ldots, X_d$ take:

$$P(Y = b \mid X_1 = a, X_2, \ldots, X_d) = 0$$

What now???

See you next week!