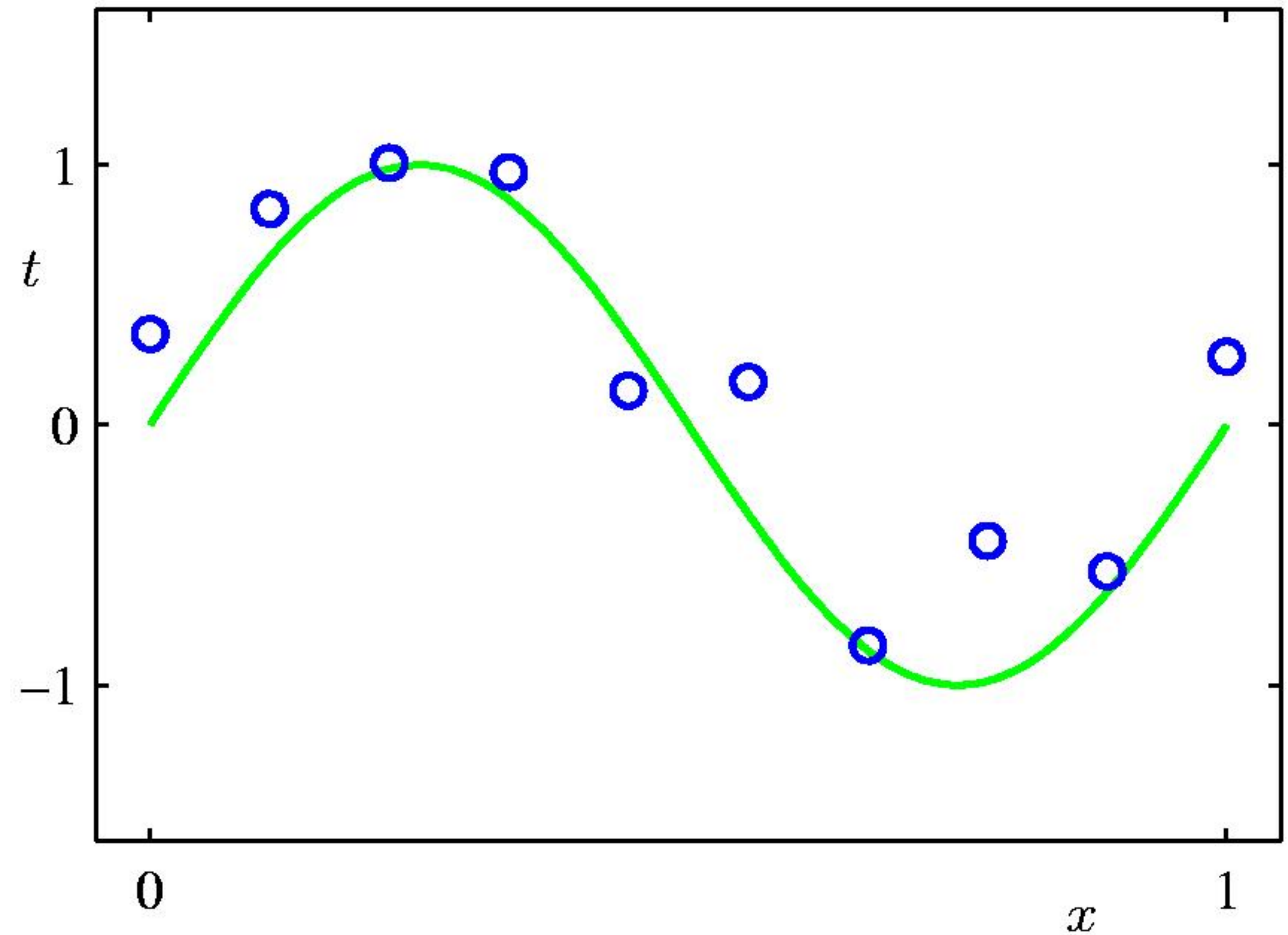
CSE 847: Machine Learning

Probability Basics

Jiayu Zhou
Computer Science & Engineering
Michigan State University

Noise



Noise

http://mrme.me/sharpening-and-noise-removal/

Probability Theory

- Uncertainty arises both through noise on measurements, as well as through the finite size of data sets.

- Probability theory provides a consistent framework for the quantification and manipulation of uncertainty and forms one of the central foundations for machine learning.

- When combined with decision theory, it allows us to make optimal predictions given all the information available to us, even though that information may be incomplete or ambiguous.

Outline

- Basic concepts in probability theory
- Bayes' rule
- Random variable and distributions

Definition of Probability

Experiment: toss a coin twice

Sample space: possible outcomes of an experiment

S = {HH, HT, TH, TT}

Event: a subset of possible outcomes

A={HH}, B={HT, TH}

Probability of an event : an number assigned to an event p(A)

Axiom 1: p(A) ≥ 0

Axiom 2: p(S) = 1

Axiom 3: For every sequence of disjoint events

$$\Pr\left(\bigcup_i A_i\right) = \sum_i \Pr(A_i)$$

Example: p(A) = n(A)/N: frequentist statistics

Joint Probability

- For events A and B, **joint probability** P(AB) stands for the probability that both events happen.
  - AB (or A ∩ B) => simultaneous occur. of events A and B
- Example:
  - A={HH, HT}, B={HH, TH}. What is P(AB)?
  - A={HH}, B={HT, TH}. What is P(AB)?

Independence

- Two events ***A and B are independent*** in case

$$p(\mathrm{AB}) = p(\mathrm{A})p(\mathrm{B})$$

- Can be extended to multiple events

$$p(\cap_i \mathrm{A}_i) = \prod_i p(\mathrm{A}_i)$$

Independence (cont.)

- Consider the experiment of tossing a coin twice
- Example I:
  - A = {HT, HH}, B = {HT}
  - Will event A independent from event B?
- Example II:
  - A = {HT}, B = {TH}
  - Will event A independent from event B?
- Disjoint ≠ Independence

- If A is independent from B, B is independent from C, will A be independent from C?

Conditioning

- If A and B are events with Pr(A) > 0, the ***conditional probability of B given A*** is

$$\Pr(B \mid A) = \frac{\Pr(AB)}{\Pr(A)}$$

Conditioning

- If A and B are events with P (A) > 0, the ***conditional probability of B given A*** is

$$\Pr(B \mid A) = \frac{\Pr(AB)}{\Pr(A)}$$

	Women	**Men**
Success	200	1800
Failure	1800	200

A = {Patient is a Women}

B = {Drug fails}

p (B|A) = ?

p (A|B) = ?

Conditioning

- If A and B are events with P (A) > 0, the ***conditional probability of B given A*** is

$$\Pr(B \mid A) = \frac{\Pr(AB)}{\Pr(A)}$$

	Women	Men
Success	200	1800
Failure	1800	200

A = {Patient is a Women}

B = {Drug fails}

p (B|A) = ?

p (A|B) = ?

Given A is independent from B, what is the relationship between p(A|B) and p(A)?

Which Drug is Better?

	Women		Men	
	Drug I	Drug II	Drug I	Drug II
Success	200	10	19	1000
Failure	1800	190	1	1000

Simpson's Paradox: View I

	Women		Men	
	Drug I	Drug II	Drug I	Drug II
Success	200	10	19	1000
Failure	1800	190	1	1000

	Drug I	**Drug II**
Success	219	1010
Failure	1801	1190

Simpson's Paradox: View I

	Women		Men	
	Drug I	Drug II	Drug I	Drug II
Success	200	10	19	1000
Failure	1800	190	1	1000

	Drug I	**Drug II**
Success	219	1010
Failure	1801	1190

A = {Using Drug I}

B = {Using Drug II}

C = {Drug succeeds}

$p(C|A) \sim 10\%$

$p(C|B) \sim 50\%$

Drug II is better than Drug I

Simpson's Paradox: View II

	Women		Men	
	Drug I	Drug II	Drug I	Drug II
Success	200	10	19	1000
Failure	1800	190	1	1000

Looking into male and female patients individually.

- What are $p(C|A)$ and $p(C|B)$ for female patients?
- What are $p(C|A)$ and $p(C|B)$ for male patients?

Simpson's Paradox: View II

	Women		Men	
	Drug I	Drug II	Drug I	Drug II
Success	200	10	19	1000
Failure	1800	190	1	1000

	Women	
	Drug I	Drug II
Success	200	10
Failure	1800	190

Female Patient

A = {Using Drug I}

B = {Using Drug II}

C = {Drug succeeds}

Pr(C|A) ~ 20%

Pr(C|B) ~ 5%

Simpson's Paradox: View II

	Women		Men	
	Drug I	Drug II	Drug I	Drug II
Success	200	10	19	1000
Failure	1800	190	1	1000

	Men	
	Drug I	Drug II
Success	19	1000
Failure	1	1000

Male Patient

Pr(C|A) ~ 100%

Pr(C|B) ~ 50%

Simpson's Paradox: View II

	Women		Men	
	Drug I	Drug II	Drug I	Drug II
Success	200	10	19	1000
Failure	1800	190	1	1000

Drug I is better than Drug II

	Female Patient	Male Patient
A = {Using Drug I}		
B = {Using Drug II}	Pr(C|A) ~ 20%	Pr(C|A) ~ 100%
C = {Drug succeeds}	Pr(C|B) ~ 5%	Pr(C|B) ~ 50%

Conditional Independence

- Event A and B are ***conditionally independent given C*** in case

  $p(AB|C)=p(A|C)p(B|C)$

- A set of events $\{A_i\}$ is conditionally independent given C in case

$$\Pr(\bigcup_i A_i \mid C) = \prod_i \Pr(A_i \mid C)$$

Conditional Independence (cont'd)

- Example: There are three events: A, B, C
  - $p(A) = p(B) = p(C) = 1/5$
  - $p(A,C) = p(B,C) = 1/25$, $p(A,B) = 1/10$
  - $p(A,B,C) = 1/125$
  - Whether A, B are independent?
  - Whether A, B are conditionally independent given C?

Conditional Independence (cont'd)

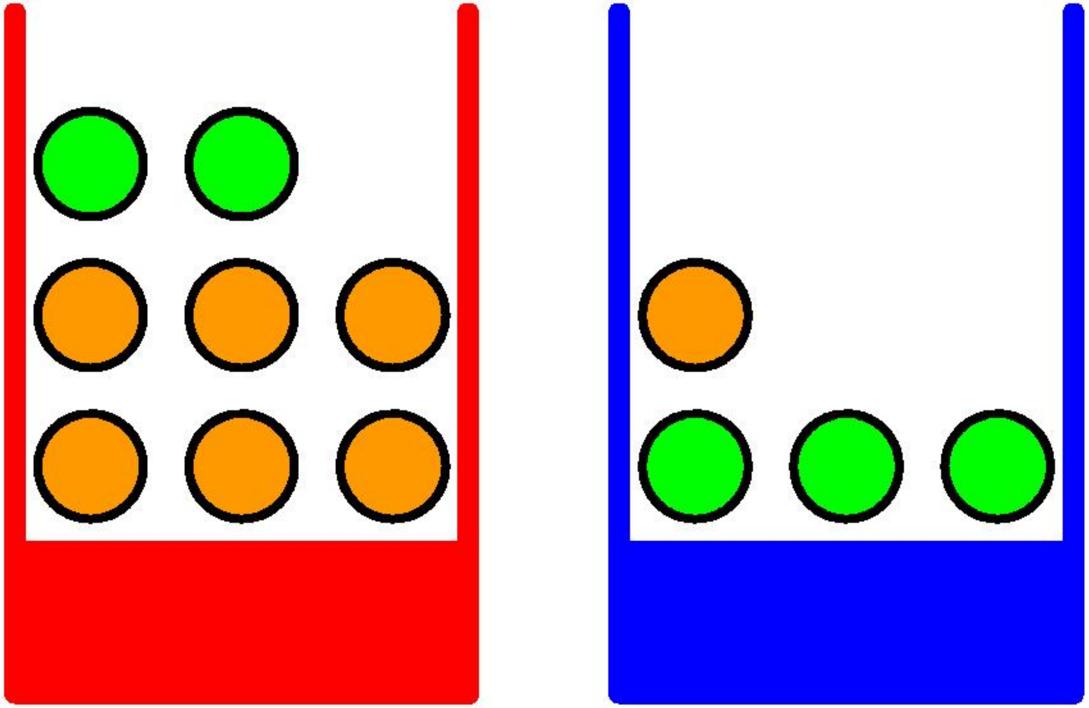
- Example: There are three events: A, B, C
  - $p(A) = p(B) = p(C) = 1/5$
  - $p(A,C) = p(B,C) = 1/25$, $p(A,B) = 1/10$
  - $p(A,B,C) = 1/125$
  - Whether A, B are independent?
  - Whether A, B are conditionally independent given C?
- *A and B are independent* $\neq$ *A and B are conditionally independent*

Outline

- Basic concepts in probability theory
- Bayes’ rule
- Random variable and distributions

A simple Example

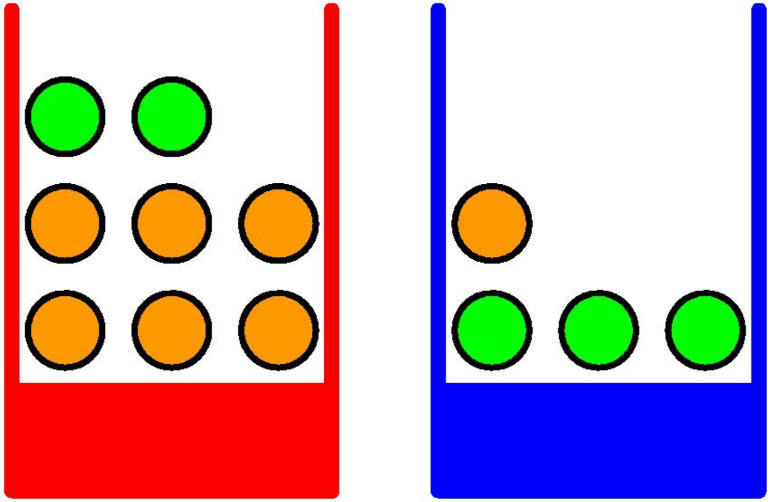
The **probability** of an event is the fraction of times that event occurs out of the total number of trials, in the limit that the total number of trials goes to infinity.



Apples and Oranges

P(B=r) = 40%, P(B=b) = 60%

A Simple Example

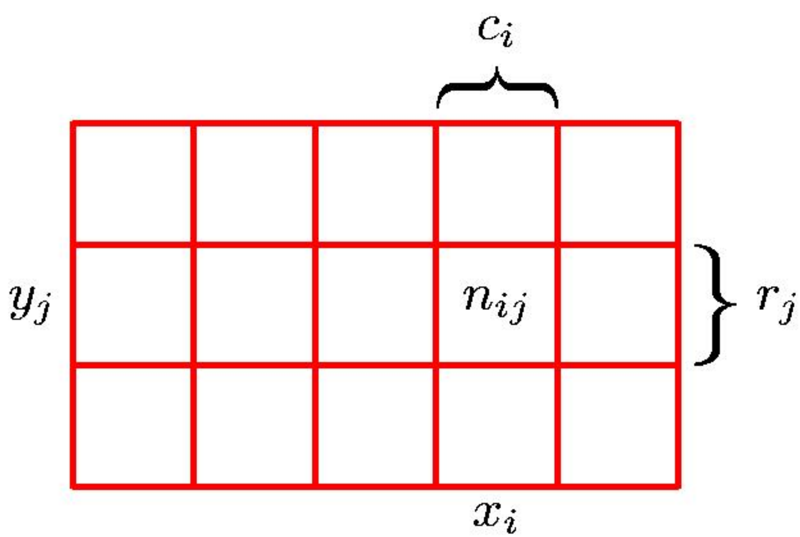


		F	
		a	o
B	r	10	30
	b	45	15

Apples and Oranges

P(B=r) = 40%, P(B=b) = 60%

Probability Theory



Marginal Probability

$$p(X = x_i) = \frac{c_i}{N}.$$

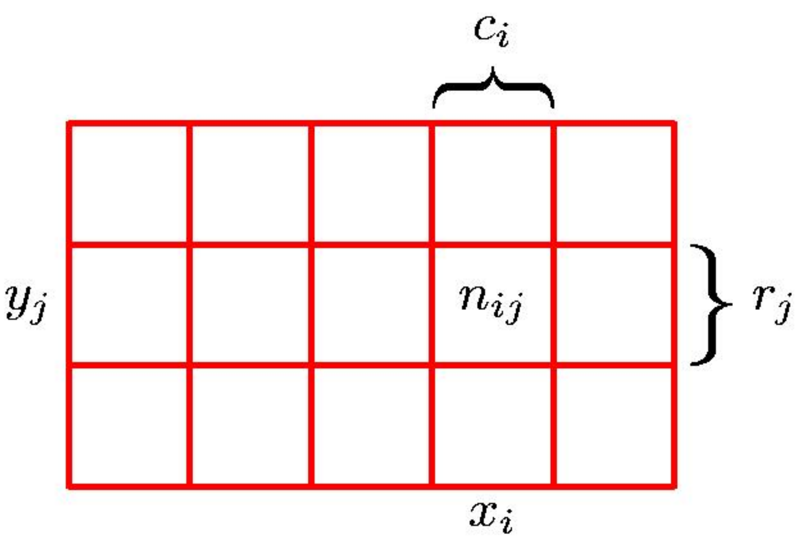
Joint Probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Conditional Probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

Probability Theory



Sum Rule

$$p(X = x_i) = \frac{c_i}{N} = \frac{1}{N}\sum_{j=1}^{L} n_{ij}$$

$$= \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

Product Rule

$$\begin{aligned} p(X = x_i, Y = y_j) &= \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} \\ &= p(Y = y_j | X = x_i) p(X = x_i) \end{aligned}$$

The Rules of Probability

Sum Rule $$p(X) = \sum_Y p(X, Y)$$

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

These two simple rules form the basis for all of the probabilistic machinery that we need.

Bayes' Theorem

From the product rule, together with the symmetry property *p(X, Y) = p(Y,X),* we obtain the following relationship between conditional probabilities:

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

posterior ∝ likelihood × prior

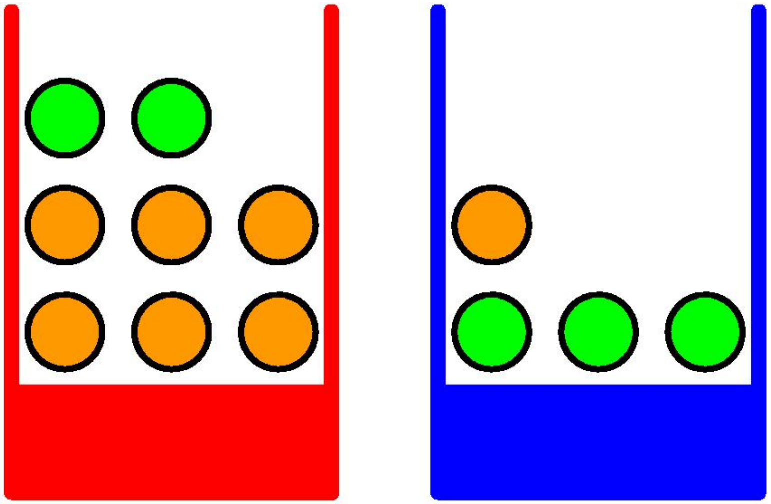
Bayes' theorem plays a central role in pattern recognition and machine learning.

$$p(X) = \sum_Y p(X|Y)p(Y)$$

normalization constant

Illustration of Bayes' Theorem

Suppose we are told that a piece of fruit has been selected and it is an orange, and we would like to know which box it came from.



$$p(B = r \mid F = o) = ?$$
$$p(B = b \mid F = o) = ?$$

Apples and Oranges

P(B=r) = 40%, P(B=b) = 60%

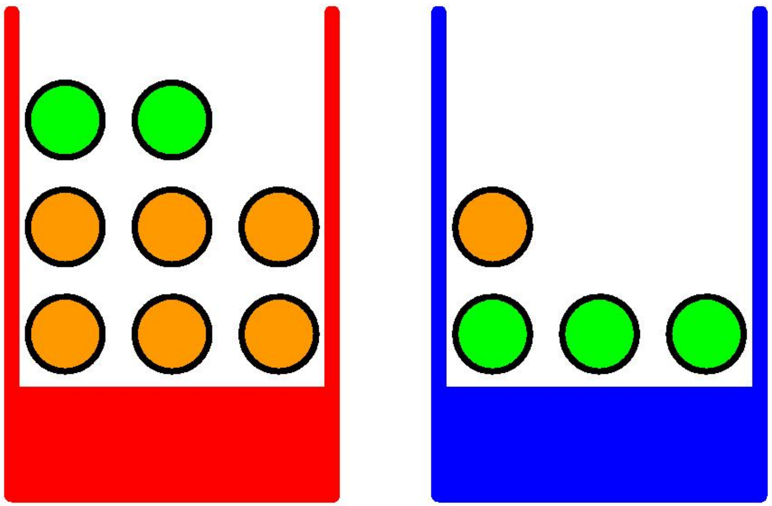
Illustration of Bayes' Theorem

Suppose we are told that a piece of fruit has been selected and it is an orange, and we would like to know which box it came from.

Apples and Oranges

$p(B = r|F = o) = p(F = o|B = r)p(B = r)/ p(F = o)$
$p(B = b|F = o) = p(F = o|B = b)p(B = b)/ p(F = o)$

$p(B = r|F = o) / p(B = b|F = o) = 2/1$
$p(B = r|F = o) + p(B = b|F = o) = 1$



$\Rightarrow$ $p(B = r|F = o) = 2/3$
$p(B = b|F = o) = 1/3$

$P(B=r) = 40\%, P(B=b) = 60\%$

Interpretation of Bayes' Theorem

$$p(B\,|F) = p(F\,|B)\; p(B) \;/\; p(F)$$

posterior ∝ likelihood × prior

- P(B): ***prior probability*** *because it is the* probability available *before we observe the identity of the fruit.*

- *p(B|F):* ***posterior probability*** *because it is the probability* obtained *after we have observed F.*

Outline

- Basic concepts in probability theory
- Bayes' rule
- Random variable and distributions

Random Variable and Distribution

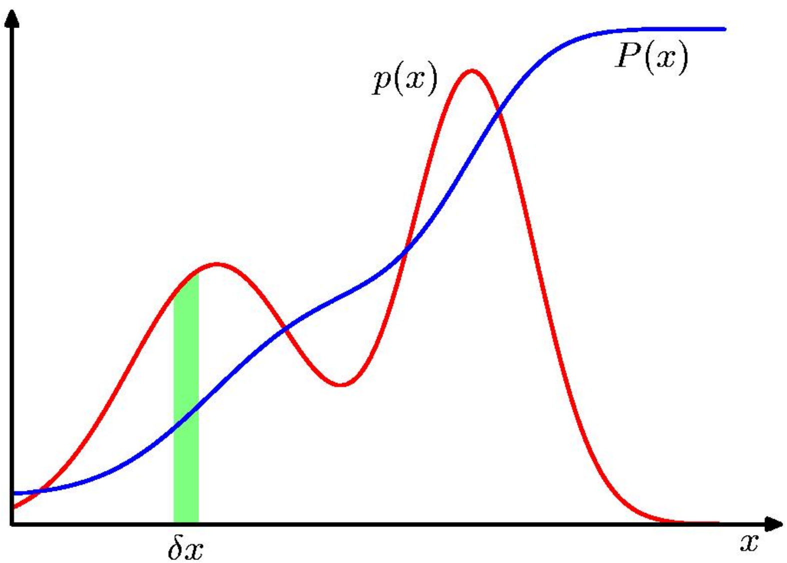
- A ***random variable X*** is a numerical outcome of a random experiment
- The ***distribution*** of a random variable is the collection of possible outcomes along with their probabilities:
  - Discrete case: $\Pr(X = x) = p_\theta(x)$
  - Continuous case: $\Pr(a \le X \le b) = \int_a^b p_\theta(x)dx$

Random Variable: Example

- Let S be the set of all sequences of three rolls of a die. Let X be the **sum of the number of dots on the three rolls**.
- What are the possible values for X?
- $\Pr(X = 5) = ?$, $\Pr(X = 10) = ?$

Probability Density Function

❑**Probability Density**: If the probability of a real-valued variable x falling in the interval (x, x + δx) is given by p(x)δx for δx → 0, then p(x) is called the probability density over x.



The probability density p(x) must satisfy the two conditions:

$$p(x) \geqslant 0 \qquad \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

$$p(x \in (a,b)) = \int_a^b p(x)\,\mathrm{d}x$$

The probability density can be expressed as the derivative of a **cumulative distribution function**:

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$

Expectation

- For a random variable X~Pr(X=x), its **expectation** is

$$E[X] = \sum_x x \Pr(X = x)$$

  - In an empirical sample, x1, x2,…, xN, $E[X] = \frac{1}{N}\sum_{i=1}^{N} x_i$

- Continuous case: $E[X] = \int_{-\infty}^{\infty} x p_\theta(x) dx$

- Expectation of sum of random variables

$$E[X_1 + X_2] = E[X_1] + E[X_2]$$

Expectation (cont.)

The average value of some function *f(x) under a* probability distribution *p(x) is called the* ***expectation*** *of f(x):*

Discrete

$$\mathbb{E}[f] = \sum_x p(x) f(x)$$

Approximate Expectation

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

Continuous

$$\mathbb{E}[f] = \int p(x) f(x) \, \mathrm{d}x$$

Variances

The **variance** of f(x) denoted as var[f] provides a measure of how much variability there is in f(x) around its mean value E[f(x)].

$$\operatorname{var}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

It can be expressed as

$$\operatorname{var}[f] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

Covariances

For two random variables x and y, the **covariance** is defined by

$$\begin{aligned}\mathrm{cov}[x, y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]\end{aligned}$$

It expresses the extent to which x and y vary together. If x and y are independent, then their covariance vanishes.

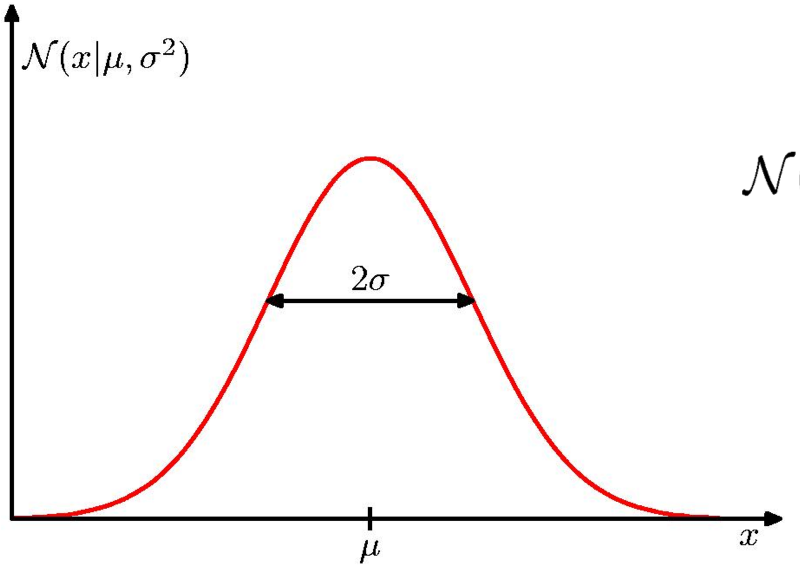
The **covariance** between two vectors of random variables x and y is a matrix

$$\begin{aligned}\mathrm{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\{\mathbf{y}^{\mathrm{T}} - \mathbb{E}[\mathbf{y}^{\mathrm{T}}]\}\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}].\end{aligned}$$

The Gaussian Distribution

The normal or Gaussian distribution is one of the most important probability distributions for continuous variables.

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



$$\mathcal{N}(x|\mu,\sigma^2) > 0 \qquad \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right)\,\mathrm{d}x = 1$$

The square root of the variance, given by σ, is called the **standard deviation**, and the reciprocal of the variance is called the **precision**.

Gaussian Mean and Variance

The average value of x is given by

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x \, \mathrm{d}x = \mu$$

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x^2 \, \mathrm{d}x = \mu^2 + \sigma^2$$

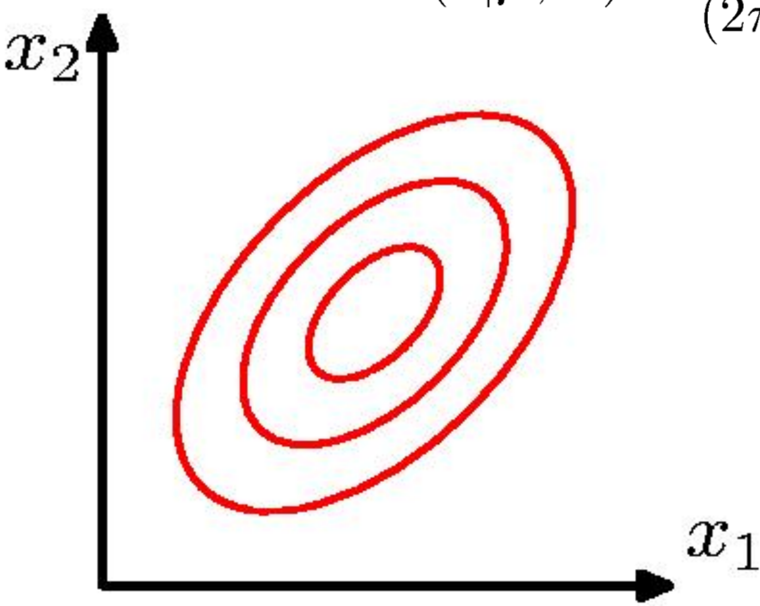
The variance of x is given by

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

The Multivariate Gaussian

The Gaussian distribution defined over a D-dimensional vector x of continuous variables is given by

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

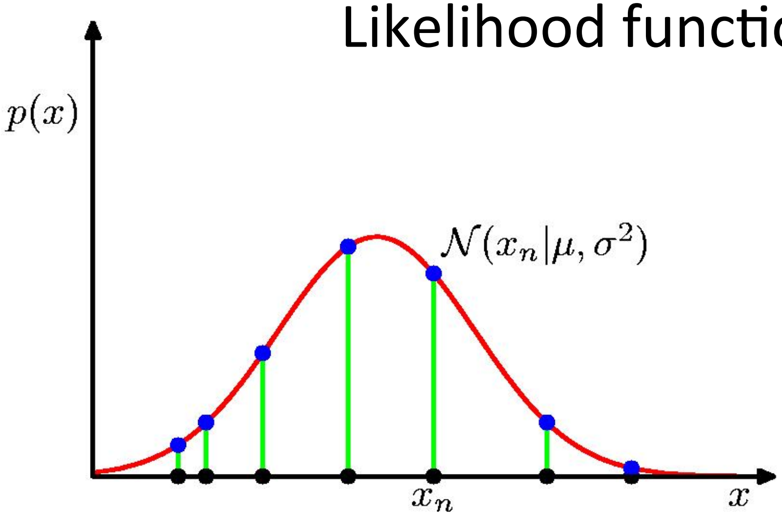


Σ is the covariance, which is a DxD matrix.

Gaussian Parameter Estimation

We are given a data set of N observations **x**={x_1, x_2, …, x_n} of the scalar variable x. We shall suppose that the observations are drawn independently from a Gaussian distribution whose mean and variance are unknown, and need to be estimated.

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



Maximum Likelihood:
Determine values for the unknown parameters in the Gaussian by maximizing the likelihood function.

Maximum (Log) Likelihood

It is more convenient to maximize the **log of the likelihood** function:

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

$$\Rightarrow \left\{ \begin{array}{l} \mu_{\mathrm{ML}} = \dfrac{1}{N}\displaystyle\sum_{n=1}^{N} x_n \\[2em] \sigma^2_{\mathrm{ML}} = \dfrac{1}{N}\displaystyle\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2 \end{array} \right.$$

μ_{ML}: Sample mean

σ^2_{ML}: Sample variance

MAP: A Step towards Bayes

We take a step towards a more Bayesian approach and introduce a prior distribution over the parameters.

MAP (maximum posterior): Determine the parameters by finding the most probable values given the data, in other words by maximizing the posterior distribution.

Bayes' theorem: posterior $\propto$ likelihood $\times$ prior

Full Bayesian Approach

In MAP, we are still making a *point estimate* and so this does not yet amount to a Bayesian treatment.

In a fully Bayesian approach, we should integrate over all values of the parameter (marginalization).

Decision Theory

- Suppose we have an input vector **x** together with a corresponding vector t of target variables, and our goal is to predict t given a new value for **x**.
  - *Regression*: t comprises continuous variables
  - *Classification*: t represents class labels

Decision Theory

- Inference step

  Determine either $p(\mathbf{x}, t)$ or $p(t|\mathbf{x})$. It gives us the most complete probabilistic description of the situation.

- Decision step (how to make optimal decisions)

  For given x, determine optimal t.

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}$$

posterior prior

Inference and Decision (1)

Three approaches to solving decision problems:

❑First solve the inference problem of determining the class-conditional densities $p(x|C_k)$ for each class C_k individually. Also separately infer the prior class probabilities $p(C_k)$. Then use Bayes' theorem to find the posterior probabilities in the form:

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}$$

Generative model: Equivalently, we can model the joint distribution *$p(x, C_k)$ directly* and then normalize to obtain the posterior probabilities.

Inference and Decision (2)

- First solve the inference problem of determining the posterior class probabilities $p(C_k|x)$, and then subsequently use decision theory to assign each new x to one of the classes. Approaches that model the posterior probabilities directly are called **discriminative models**.

- Find a function $f(x)$, called a **discriminant function**, which maps each input x directly onto a class label. For instance, in the case of two-class problems, $f(\cdot)$ might be binary valued and such that $f = 0$ represents class C_1 and $f = 1$ represents class C_2. In this case, *probabilities play no role*.

Next Class

- Topic
  - Linear Algebra Basics

- Reading
  - Book Ch. 1, 2