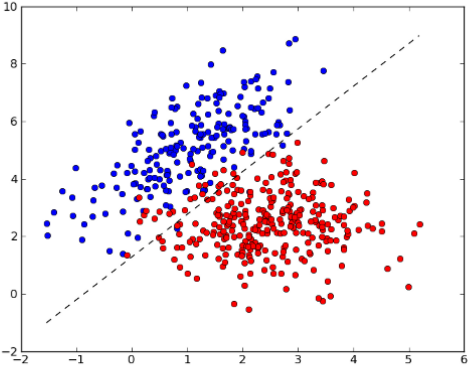
Probability Basics

Jiayu Zhou

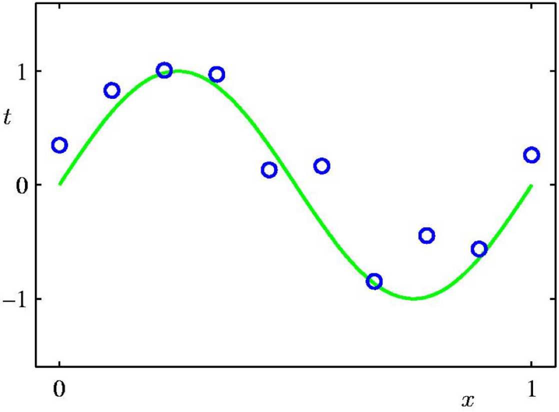
[1]Department of Computer Science and Engineering
Michigan State University
East Lansing, MI USA

Table of contents

Noise



Noise



Noise in Sensors

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.

Probability in Machine Learning

x is a data point (vector of features) and y is a label we would like to predict.

Probability in Machine Learning

x is a data point (vector of features) and y is a label we would like to predict.

- $p(y)$: prior, e.g., news classification during election, fairness

Probability in Machine Learning

x is a data point (vector of features) and y is a label we would like to predict.

- $p(y)$: prior, e.g., news classification during election, fairness
- $p(y|x)$: posterior, confidence of prediction

Probability in Machine Learning

x is a data point (vector of features) and y is a label we would like to predict.

- $p(y)$: prior, e.g., news classification during election, fairness
- $p(y|x)$: posterior, confidence of prediction
- $p(x)$: generative models, e.g., Generative Adversarial Networks

Probability in Machine Learning

x is a data point (vector of features) and y is a label we would like to predict.

- $p(y)$: prior, e.g., news classification during election, fairness
- $p(y|x)$: posterior, confidence of prediction
- $p(x)$: generative models, e.g., Generative Adversarial Networks
- $p(x|y)$: class conditioned density, e.g., conditional generative models

Definition of Probability

- **Experiment**: toss a coin twice
- **Sample space**: possible outcomes of an experiemnt
  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: $p(A) \geq 0$
  - Axiom 2: $p(S) = 1$
  - Axiom 3: For every sequence of *disjoint* events

$$p(\cup_i A_i) = \sum_i p(A_i)$$

  - 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 $A \cap B$) $\Rightarrow$ simultaneous occur. of events A and B

Joint Probability

- For events A and B, **joint probability** $p(AB)$ stands for the probability that *both* events happen.
  - AB (or $A \cap B$) $\Rightarrow$ 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

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

- Can be extended to multiple events

$$p(\cap_i A_i) = \prod_i p(A_i)$$

Independence (cont.)

- Consider the experiment of tossing a coin twice
  - Example I: $A = \{HT, HH\}, B = \{HT\}$. Will event A independent from event B?

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?

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 $\neq$ Independence.

Conditioning

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

$$p(B|A) = \frac{p(AB)}{p(A)}.$$

Conditioning

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

$$p(B|A) = \frac{p(AB)}{p(A)}.$$

- Example

	Women	Men
Success	200	1800
Failure	1800	200

A = Patient is a Woman, B = Drug fails, what are $p(B|A)$ and $p(A|B)$?

Conditioning

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

$$p(B|A) = \frac{p(AB)}{p(A)}.$$

- Example

	Women	Men
Success	200	1800
Failure	1800	200

A = Patient is a Woman, B = Drug fails, what are $p(B|A)$ and $p(A|B)$?

- 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

Which Drug is Better?

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

- View I: Comparing $p(C|A)$ and $p(C|B)$, where C={Drug succeed}, and A={Using Drug I}, and B={Using Drug II}.

	Drug I	Drug II
Success	219	1010
Failure	1801	1190

Which Drug is Better?

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

- View I: Comparing $p(C|A)$ and $p(C|B)$, where C={Drug succeed}, and A={Using Drug I}, and B={Using Drug II}.

	Drug I	Drug II
Success	219	1010
Failure	1801	1190

- $p(C|A) \approx 10\%$ and $p(C|B) \approx 50\%$.
- **Drug II is better than Drug I.**

Which Drug is Better?

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

- View II: Looking into male and female patients individually. What are $p(C|A)$ and $p(C|B)$ for female and male patients respectively?

Women	Drug I	Drug II
Success	200	10
Failure	1800	190

$p(C|A) = 10\%$, $p(C|B) = 5\%$

Which Drug is Better?

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

- View II: Looking into male and female patients individually. What are $p(C|A)$ and $p(C|B)$ for female and male patients respectively?

Women	Drug I	Drug II
Success	200	10
Failure	1800	190

$p(C|A) = 10\%,\ p(C|B) = 5\%$

Men	Drug I	Drug II
Success	19	1000
Failure	1	1000

$p(C|A) \approx 100\%,\ p(C|B) = 50\%$

Which Drug is Better?

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

- View II: Looking into male and female patients individually. What are $p(C|A)$ and $p(C|B)$ for female and male patients respectively?

Women	Drug I	Drug II
Success	200	10
Failure	1800	190

$p(C|A) = 10\%,\ p(C|B) = 5\%$

Men	Drug I	Drug II
Success	19	1000
Failure	1	1000

$p(C|A) \approx 100\%,\ p(C|B) = 50\%$

Drug I is better than Drug II

Conditional Independence

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

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

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

$$p(\cup_i A_i|C) = \prod_i p(A_i|C)$$

Conditional Independence (cont'd)

Example: There are three events A, B, C

- $p(A) = p(B) = p(C) = 1/5$
- $p(A, C) = p(B, C) = 1/25, p(A, B) = 1/10$
- $p(A, B, C) = 1/125$

Conditional Independence (cont'd)

Example: There are three events A, B, C

- $p(A) = p(B) = p(C) = 1/5$
- $p(A, C) = p(B, C) = 1/25, p(A, B) = 1/10$
- $p(A, B, C) = 1/125$

Answer the following question:

- Whether A, B are independent?

Conditional Independence (cont'd)

Example: There are three events A, B, C

- $p(A) = p(B) = p(C) = 1/5$
- $p(A, C) = p(B, C) = 1/25, p(A, B) = 1/10$
- $p(A, B, C) = 1/125$

Answer the following question:

- 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

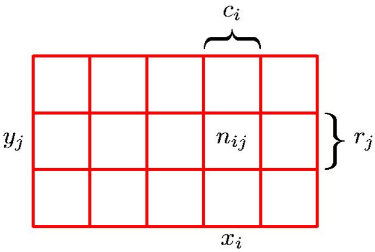
- $p(A) = p(B) = p(C) = 1/5$
- $p(A, C) = p(B, C) = 1/25, p(A, B) = 1/10$
- $p(A, B, C) = 1/125$

Answer the following question:

- 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.

Probability Computation



Marginal Probability

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

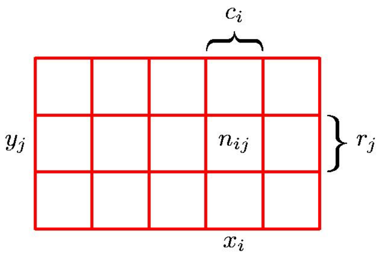
Conditional Probability

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

Joint Probability

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

Probability Computation



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

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

$$= p(Y = y_i | X = x_i)p(X = x_i)$$

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:

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

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

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:

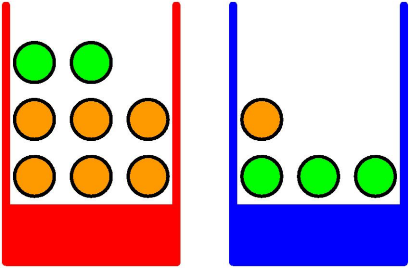
$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

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

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

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

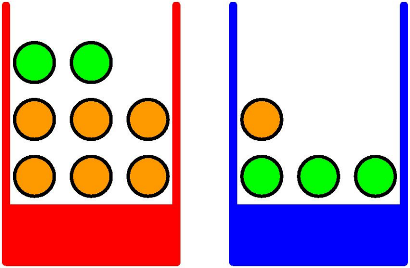
Illustration of Bayes' Theorem



apples and oranges, $p(\text{Box} = \text{red}) = 40\%$, $p(\text{Box} = \text{blue}) = 60\%$.

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.

Illustration of Bayes' Theorem

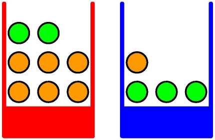


apples and oranges, $p(\text{Box} = \text{red}) = 40\%$, $p(\text{Box} = \text{blue}) = 60\%$.

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.

$$p(B = r | F = o) = ?, p(B = b | F = o) = ?$$

Illustration of Bayes' Theorem



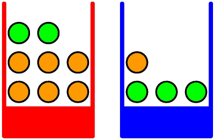
apples and oranges,
$p(\text{Box} = \text{red}) = 40\%$,

$p(\text{Box} = \text{blue}) = 60\%$.

$$p(B = r | F = o) = \frac{p(F = o | B = r) p(B = r)}{p(F = o)}$$

$$p(B = b | F = o) = \frac{p(F = o | B = b) p(B = b)}{p(F = o)}$$

Illustration of Bayes' Theorem



apples and oranges,
$p(\text{Box} = \text{red}) = 40\%$,
$p(\text{Box} = \text{blue}) = 60\%$.

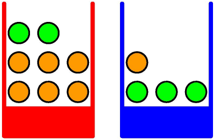
$$p(B = r|F = o) = \frac{p(F = o|B = r)p(B = r)}{p(F = o)}$$

$$p(B = b|F = o) = \frac{p(F = o|B = b)p(B = b)}{p(F = o)}$$

We have

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

Illustration of Bayes' Theorem



apples and oranges,
$p(\text{Box} = \text{red}) = 40\%$,
$p(\text{Box} = \text{blue}) = 60\%$.

$$p(B = r|F = o) = \frac{p(F = o|B = r)p(B = r)}{p(F = o)}$$

$$p(B = b|F = o) = \frac{p(F = o|B = b)p(B = b)}{p(F = o)}$$

We have

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

Therefore

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

Interpretation of Bayes' Theorem

$$p(B|F) = \frac{p(F|B)p(B)}{p(F)}$$

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

- $p(B)$: 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.

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: $p(X = x) = p(x)$
  - Continuous case: $p(a \leq X \leq b) = \int_a^b p(x)dx$

Expectation

- For a random variable $X \sim p(X = x)$, its **expectation** is

$$\mathbb{E}(X) = \sum_x xp(X = x)$$

  - In an empirical sample, $x_1, x_2, \ldots, x_N$, $\mathbb{E}[X] = \frac{1}{N}\sum_{i=1}^N x_i$
- Continuous case: $E[X] = \int_{-\infty}^{\infty} xp(x)dx$
- Expectation of sum of random variables

$$\mathbb{E}[X_1 + X_2] = \mathbb{E}[X_1] + \mathbb{E}[X_2]$$

Expectation of a function

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

- Discrete $\mathbb{E} = \sum_x p(x)f(x)$
  Approximate Expectation $\mathbb{E} \approx \frac{1}{N}\sum_{n=1}^{N} f(x_n)$
- Continuous $\mathbb{E}[f] = \int p(x)f(x)dx$

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 $\mathbb{E}[f(x)]$.

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

Covariances

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

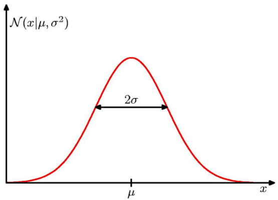
$$
\begin{aligned}
\mathrm{cov}[x, y] &= \mathbb{E}_{x,y}[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}] \\
&= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]
\end{aligned}
$$

- 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

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

The Gaussian Distribution

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



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

- The square root of the variance, given by σ, is called the standard deviation, and the reciprocal of the variance is called the precision.
- $\mathcal{N}(x|\mu,\sigma^2) > 0$ and $\int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2)dx = 1$

Gaussian Mean and Variance

- The average value of x is given by

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

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

- The variance of x is given by

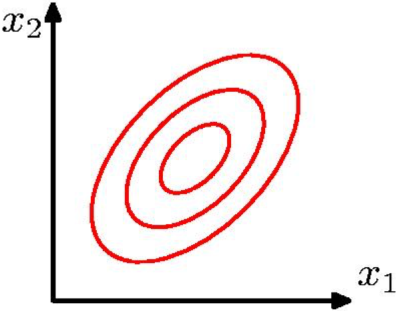
$$\text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

The Multivariate Gaussian

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

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

where $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ is the covariance



Gaussian Parameter Estimation

- We are given a data set of N observations $\mathrm{x} = \{x_1, x_2, \ldots, x_n\}$ 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.

Gaussian Parameter Estimation

- We are given a data set of N observations $\mathrm{x} = \{x_1, x_2, \ldots, x_n\}$ 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 $p(\mathrm{x}|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2)$
- **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:

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

Maximum (Log) Likelihood

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

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

- Therefore we have

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n \qquad \text{Sample mean}$$

$$\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2 \qquad \text{Sample variance}$$

Maximum a Posterior

- 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.
-

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

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
- **Inference Step**: Determine either $p(\mathrm{x}, t)$ or $p(t|\mathrm{x})$. It gives us the most complete probabilistic description of the situation.
- **Decision Step**: How to make optimal decision.
- Three approaches.

Inference and Decision: (1)

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

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

Inference and Decision: (1)

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

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

- Generative model: Equivalently, we can model the joint distribution $p(\mathrm{x}, \mathcal{C}_k)$ directly and then normalize to obtain the posterior probabilities.

Inference and Decision: (2)

- First solve the inference problem of determining the posterior class probabilities $p(\mathcal{C}_k|\mathrm{x})$, 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.

Inference and Decision: (2)

- First solve the inference problem of determining the posterior class probabilities $p(\mathcal{C}_k|\mathrm{x})$, 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 $f(\mathrm{x})$, 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 $\mathcal{C}_1$ and $f = 1$ represents class $\mathcal{C}_2$. In this case, probabilities play no role.

Next Class

- Topic
  - Linear Algebra Basics
- Reading
  - Book Ch. 1,2