CS 1675 Introduction to Machine Learning
Lecture 5

Density estimation

Milos Hauskrecht
milos@pitt.edu
5329 Sennott Square

Review of probabilities

Probability theory

Studies and describes random processes and their outcomes

- **Random processes may result in multiple different outcomes**
- **Example 1: coin flip**
  – Outcome is either head or tail (binary outcome)
  – Fair coin: outcomes are equally likely

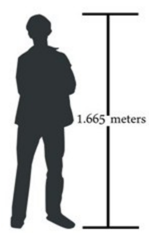
- **Example 2: sum of numbers obtained by rolling 2 dice**
  – Outcome number in between 2 to 12
  – Fair dices: outcome 2 is less likely then 3

Probability theory

Studies and describes random processes and their outcomes

- **Random processes may have multiple different outcomes**

- **Example 3: height of a person**
  – Select randomly a person from your school/city and report her height
  – Outcomes can be real numbers



- **And many others related to measurements, lotteries, etc**

Probabilities

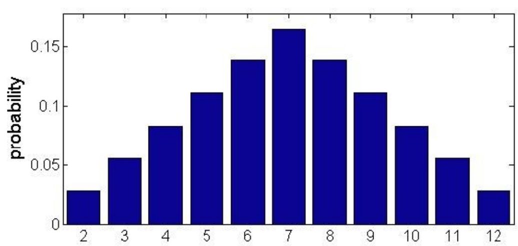
When the process is repeated many times outcomes occur with certain relative frequencies or **probabilities**

- **Example 1: coin flip**
  - **Fair coin:** outcomes are equally likely
    - Probability of head is 0.5 and tail is 0.5
  - Biased coin
    - Probability of head is 0.8 and tail is 0.2
    - Head outcome is 4 times more likely than tail

Probabilities

When the process is repeated many times outcomes occur with certain relative frequencies or **probabilities**

- **Example 2: sum of numbers obtained by rolling 2 dice**
  - Outcome number in between 2 to 12
  - Fair dice: outcome 2 is less likely then 3
    4 is less likely then 3, etc



Probability distribution function

Discrete (mutually exclusive) outcomes – the chance of outcomes is represented by **a probability distribution function**

- **probability distribution function – assigns a number between 0 and 1 to every outcome**
- **Example 1: coin flip**
  - Biased coin
    - Probability of head is 0.8 and tail is 0.2
    - Head outcome is 4 time more likely than tail

P(tail) = 0.2
P(head) = 0.8

$$P(coin) = \begin{bmatrix} 0.2 \\ 0.8 \end{bmatrix}$$

- **What is the condition we need to satisfy ?**

Probability distribution function

Discrete (mutually exclusive) outcomes – the chance of outcomes is represented by **a probability distribution function**

- **probability distribution function – assigns a number between 0 and 1 to every outcome**
- **Example 1: coin flip**
  - Biased coin
    - Probability of head is 0.8 and tail is 0.2
    - Head outcome is 4 time more likely than tail

P(tail) = 0.2
P(head) = 0.8

$$P(coin) = \begin{bmatrix} 0.2 \\ 0.8 \end{bmatrix}$$

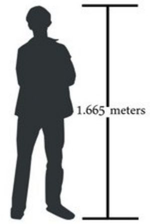
- **What is the condition we need to satisfy ?**

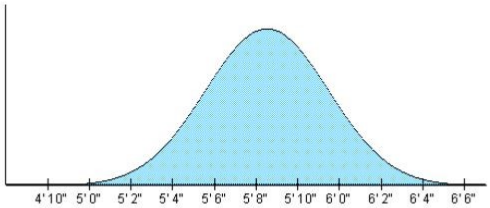
- **Sum of probabilities for discrete set of outcomes is 1**

Probability for real-valued outcomes

When the process is repeated many times outcomes occur with certain relative frequencies or **probabilities**

- **Example 3: height of a person**
  - Select randomly a person from your school/city and report her height
  - Outcomes can be real numbers
  - Different outcomes can be more or less likely



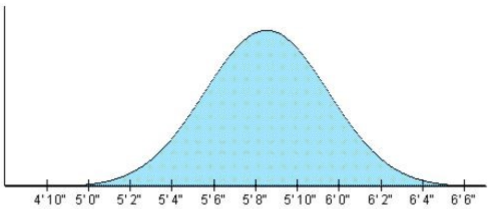


Normal (Gaussian) density

Probability density function

Real-valued outcomes – the chance of outcomes is represented by **a probability density function**

- **probability density function – p(x)**

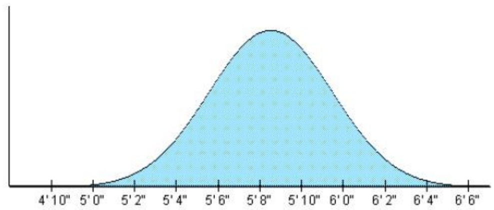


- **Condition on p(x) and 1?**

Probability density function

Real-valued outcomes – the chance of outcomes is represented by **a probability density function**

- **probability density function – p(x)**



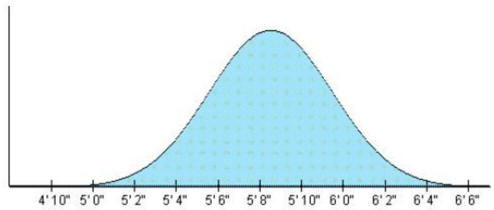
- **Conditions on p(x) and 1?**

$$\boxed{\int p(x)dx = 1}$$

Probability density function

Real-valued outcomes – the chance of outcomes is represented by **a probability density function**

- **probability density function – p(x)**

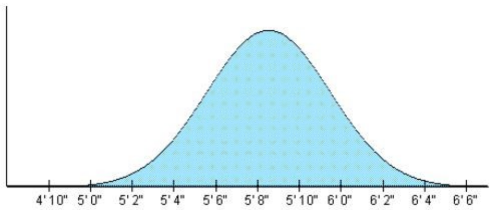


- **Can p(x) values for some x be negatives?**

Probability density function

Real-valued outcomes – the chance of outcomes is represented by **a probability density function**

- **probability density function – p(x)**

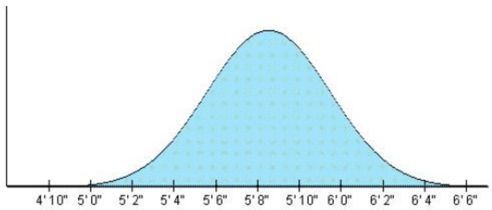


- **Can p(x) values for some x be negatives?**
- **No**

Probability density function

Real-valued outcomes – the chance of outcomes is represented by **a probability density function**

- **probability density function – p(x)**

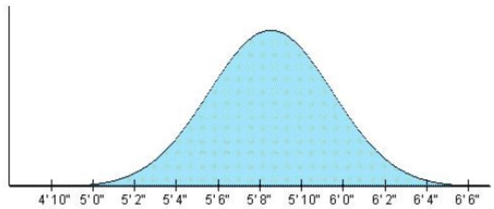


- **Can p(x) values for some x be > 1?**
  - **Remember we need** $\int p(x)dx = 1$

Probability density function

Real-valued outcomes – the chance of outcomes is represented by **a probability density function**

- **probability density function – p(x)**



- **Can p(x) values for some x be > 1?**
- **Remember we need:** $\int p(x)dx = 1$
- **Yes**

Random variable

Random variable = A function that maps observed outcomes (quantities) to real valued outcomes

Binary random variables: **Two outcomes** mapped to **0,1**

Example: Coin flip. Tail mapped to 0, Head mapped to 1

Note: Only one value for each outcome: either 0 or 1

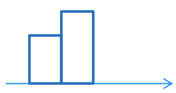
probability of tail $P(x=0)$

probability of head $P(x=1)$

Probability distribution: Assigns a probability to each possible outcome

A Biased coin

$P(\mathrm{x}) = \boxed{\begin{matrix} 0.45 \\ 0.55 \end{matrix}}$

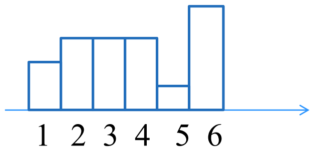


Random variable

Example: roll of a dice

– Outcomes =1,2,3,4,5,6 based on the roll of a die
– **trivial map to the same number**

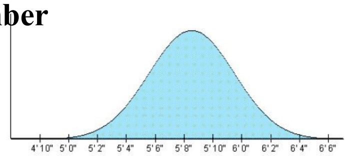
Biased dice



Example: x height of a person

Real valued outcomes

– trivial map to the same number



Probability

- **Let A be an outcome event, and ¬A its complement.**
  – **Then**

$$P(A) + P(\neg A) = ?$$

Probability

- Let A be an event, and ¬A its complement.
  - Then

$$P(A)+P(\neg A)=1$$

$$P(A\wedge\neg A)=?$$

Probability

- Let A be an event, and ¬A its complement.
  - Then

$$P(A)+P(\neg A)=1$$

$$P(A\wedge\neg A)=0$$

$$P(False)=0$$

$$P(A\vee\neg A)=?$$

Probability

- **Let A be an event, and ¬A its complement.**
  - **Then**

$$P(A) + P(\neg A) = 1$$

$$P(A \wedge \neg A) = 0$$

$$P(False) = 0$$

$$P(A \vee \neg A) = 1$$

$$P(True) = 1$$

Joint probability

Joint probability:

- **Let A and B be two events.** The probability of an event A, B occurring jointly

$$P(A \wedge B) = P(A, B)$$

We can add more events, say, A,B,C

$$P(A \wedge B \wedge C) = P(A, B, C)$$

Independence

Independence :

- Let A, B be two events. The events are independent if:

$$P(A,B) = ?$$

Independence

Independence :

- Let A, B be two events. The events are independent if:

$$P(A,B) = P(A)P(B)$$

Conditional probability

Conditional probability :

- Let A, B be two events. The conditional probability of A given B is defined as:

$$P(A \mid B) = ?$$

Conditional probability

Conditional probability :

- Let A, B be two events. The conditional probability of A given B is defined as:

$$P(A \mid B) = \frac{P(A,B)}{P(B)}$$

Product rule:

- A rewrite of the conditional probability

$$P(A,B) = P(A \mid B)P(B)$$

Bayes theorem

Bayes theorem

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Why?

$$P(A \mid B) = \frac{P(A,B)}{P(B)} \qquad P(A,B) = P(B \mid A)P(A)$$

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Density estimation

Density estimation

Density estimation: is an unsupervised learning problem

- **Goal:** Learn a model that represent the relations among attributes in the data

$$D = \{D_1, D_2, .., D_n\}$$

Data: $D_i = \mathbf{x}_i$ a vector of attribute values

Attributes:

- modeled by random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_d\}$ with
  - **Continuous or discrete valued variables**

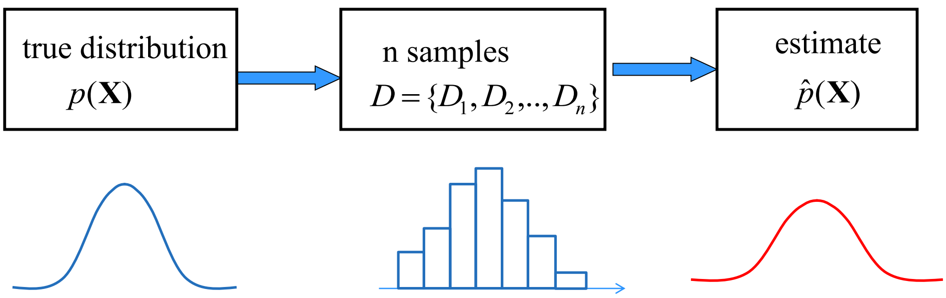
Density estimation: learn an underlying probability distribution model : $p(\mathbf{X}) = p(X_1, X_2, \ldots, X_d)$ **from D**

Density estimation

Data: $D = \{D_1, D_2, .., D_n\}$

$D_i = \mathbf{x}_i$ a vector of attribute values

Objective: estimate the model of the underlying probability distribution over variables $\mathbf{X}$, $p(\mathbf{X})$, using examples in D

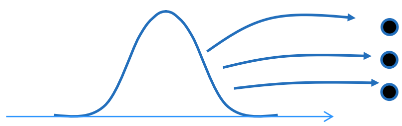


Density estimation

true distribution $p(\mathbf{X})$ → n samples $D = \{D_1, D_2, .., D_n\}$ → estimate $\hat{p}(\mathbf{X})$

Standard (iid) assumptions: Samples

- **are independent of each other**
- **come from the same (identical) distribution (fixed $p(\mathbf{X})$)**



Independently drawn instances
from the same fixed distribution

Density estimation

Types of density estimation:

Parametric

- the distribution is modeled using a set of parameters Θ

$$\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \Theta)$$

- **Example:** mean and covariances of a multivariate normal
- **Estimation:** find parameters Θ describing data D

Non-parametric

- The model of the distribution utilizes all examples in D
- As if all examples were parameters of the distribution
- **Examples:** Nearest-neighbor

Learning via parameter estimation

In this lecture we consider **parametric density estimation**

Basic settings:

- A set of random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_d\}$
- **A model of the distribution** over variables in X with parameters Θ : $\hat{p}(\mathbf{X} \mid \Theta)$

Example: Gaussian distribution with mean and variance parameters

- **Data** $D = \{D_1, D_2, .., D_n\}$

Objective: find parameters Θ such that $p(\mathbf{X} \mid \Theta)$ fits data D the best

ML Parameter estimation

Model $\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \mathbf{\Theta})$ **Data** $D = \{D_1, D_2, .., D_n\}$

- **Maximum likelihood (ML)** $\boxed{\max_{\Theta} p(D \mid \Theta, \xi)}$
  - Find Θ that maximizes likelihood $p(D \mid \Theta, \xi)$

$$P(D \mid \Theta, \xi) = P(D_1, D_2, .., D_n \mid \Theta, \xi)$$
$$= P(D_1 \mid \Theta, \xi) P(D_2 \mid \Theta, \xi) \ldots P(D_n \mid \Theta, \xi)$$
$$= \prod_{i=1}^{n} P(D_i \mid \Theta, \xi)$$

Independent examples

log-likelihood $\log p(D \mid \Theta, \xi) = \sum_{i=1}^{n} \log P(D_i \mid \Theta, \xi)$

$$\Theta_{ML} = \arg\max_{\Theta} p(D \mid \Theta, \xi) = \arg\max_{\Theta} \log p(D \mid \Theta, \xi)$$