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 than 3

• **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

\[
P(\text{tail}) = 0.2 \\
P(\text{head}) = 0.8
\]

\[
P(\text{coin}) = \begin{bmatrix} 0.2 \\ 0.8 \end{bmatrix}
\]

• 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** – \( p(x) \)

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

\[
\int p(x) \, dx = 1
\]

- **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: \( \int p(x)dx = 1 \)
- 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 with head and tail outcomes
- Tail mapped to 0, \( P(x = 0) \)
- Head mapped to 1 \( P(x = 1) \)

**Example of observed outcome sequence:**
- tail, tail, head, tail, head, head… \( \rightarrow 0, 0, 1, 0, 1, 1, … \)
Random variable

Example: roll of a dice
- Outcomes = 1, 2, 3, 4, 5, 6 based on the roll of a dice
- trivial map to the same number

Example of observed outcome sequence:
• 3, 6, 2, 6, 1, 2, 5, 4, 5, 3, 3 …

Random variable

Example: x height of a person
Real valued outcomes
- trivial map to the same number

Example of observed outcome sequence:
• 5’4”, 6’1”, 5’9”, 5’8”
Expected value of a random variable

Assume a random variable $X$ with $K$ discrete values

- Expected value of $X$ is:

$$E[X] = \sum_{i=1}^{K} p(X = x_i) x_i$$

Example: Fair dice

- Outcomes =1,2,3,4,5,6 based on the roll

Fair dice

```
1 2 3 4 5 6
```

$P = 1/6$

$$E[X] = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = 3.5$$

Expected value of a random variable

Assume a random variable $X$ with continuous values

$$E[X] = \int x \cdot p(x) dx$$

Example: $x$ height of a person

- Density function: Gaussian
- Expected value of $X$ is the center of the Gaussian distribution or its mean

```
Probability: basics

• Let A be an outcome event, and \( \neg A \) its complement.
  – Then

\[
P(A) + P(\neg A) = ?
\]

Probability: basics

• Let A be an event, and \( \neg A \) its complement.
  – Then

\[
P(A) + P(\neg A) = 1
\]

\[
P(A \land \neg A) = ?
\]
Probability: basics

• Let $A$ be an event, and $\neg A$ its complement.
  – Then

$$P(A) + P(\neg A) = 1$$
$$P(A \land \neg A) = 0$$
$$P(False) = 0$$
$$P(A \lor \neg A) = ?$$
Joint probability

**Joint probability:**
- **Let A and B be two events.** The probability of an event A, B occurring jointly
  \[ P(A \land B) = P(A, B) \]

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

\[ P(A \land B \land C) = P(A, B, C) \]

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?**

\[
P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B)}
\]
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:** \( D_i = x_i \) a vector of attribute values

**Attributes:**

- modeled by random variables \( X = \{ X_1, X_2, ..., X_d \} \) with
  - Continuous or discrete valued variables

**Density estimation:** learn an underlying probability distribution model: \( p(X) = p(X_1, X_2, ..., X_d) \) from \( D \)
Density estimation

**Data:** \[ D = \{D_1, D_2, ..., D_n\} \]
\[ D_i = x_i \quad \text{a vector of attribute values} \]

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

---

**Density estimation: iid assumptions**

**Standard (iid) assumptions:** Samples
- are independent of each other
- come from the same (identical) distribution (fixed \( p(\mathbf{X}) \))

<table>
<thead>
<tr>
<th>true distribution</th>
<th>n samples ( D = {D_1, D_2, ..., D_n} )</th>
<th>estimate ( \hat{p}(\mathbf{X}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p(\mathbf{X}) )</td>
<td>( n ) samples</td>
<td>( \hat{p}(\mathbf{X}) )</td>
</tr>
</tbody>
</table>
Density estimation

Types of density estimation:

(1) Parametric

- the distribution is modeled using a set of parameters $\Theta$
  \[ \hat{p}(X) = p(X \mid \Theta) \]
- Estimation: find parameters $\Theta$ fitting the data $D$
- Example: estimate the mean and covariance of a normal distribution
  \[ \hat{p}(x) = N(x \mid \mu, \sigma) \]

(2) Non-parametric

- The model of the distribution utilizes all examples in $D$
- As if all examples were parameters of the distribution
- Example: estimate the mean and covariance of a normal distribution
  \[ \hat{p}(X) = p(X \mid D) \]

Examples:
- histogram
- Kernel density estimation