DSC 140A

Probabilistic Modeling & Machine Learning

Lecture 12 | Part 1

Estimating Discrete Probabilities

Recall: Bayes Classifier

- The **Bayes classification rule**.

- Given a new point $\vec{x}$, predict:
  - Class 1 if $\mathbb{P}(Y = 1 \mid \vec{X} = \vec{x}) > \mathbb{P}(Y = 0 \mid \vec{X} = \vec{x})$
  - Class 0 otherwise.

- Alternative form:
  - Class 1 if
    $\mathbb{P}(\vec{X} = \vec{x} \mid Y = 1)\mathbb{P}(Y = 1) > \mathbb{P}(\vec{X} = \vec{x} \mid Y = 0)\mathbb{P}(Y = 0)$
  - Class 0 otherwise.

Bayes Error

- If $\mathbb{P}(Y = 1 \mid \vec{X} = \vec{x}) \neq 1$, there is some chance of error.
- The **Bayes classifier** achieves the lowest possible error rate.

Problem

- This assumed that we **know** the true probabilities used by Nature.
- Typically, we do not.
- But we can **estimate** them from data.

Example: Flowers

- **Example:** two species of flower (1 and 0); one species tends to have more petals than the other.
- **Goal:** given new flower with X petals, predict species, Y.
- Both X and Y are **discrete**.

Before: Joint Distribution

- Before: we somehow knew the joint distribution:

	$Y = 0$	$Y = 1$
$X = 0$	0%	0%
$X = 1$	5%	0%
$X = 2$	10%	5%
$X = 3$	15%	15%
$X = 4$	5%	20%
$X = 5$	0%	15%
$X = 6$	0%	10%

Now

- In practice, we do not know the joint distribution:

	$Y = 0$	$Y = 1$
$X = 0$	?	?
$X = 1$	?	?
$X = 2$	?	?
$X = 3$	?	?
$X = 4$	?	?
$X = 5$	?	?
$X = 6$	?	?

Data

- Suppose we observe 10 flowers.
- We can use this data to **estimate** probabilities.
- E.g., what is $\mathbb{P}(X = 4, Y = 1)$?

X	Y
5	0
3	0
4	1
4	1
2	0
5	1
2	1
5	1
4	1
3	0

Estimating Joint Probabilities

- We estimate $\mathbb{P}(X = x, Y = y)$ with:

$$\mathbb{P}(X = x, Y = y) \approx \frac{\#(X = x \text{ and } Y = y)}{n}$$

- E.g., estimate $\mathbb{P}(X = 4, Y = 1)$:
- E.g., estimate $\mathbb{P}(X = 3, Y = 0)$:
- E.g., estimate $\mathbb{P}(X = 3, Y = 1)$:

X	Y
5	0
3	0
4	1
4	1
2	0
5	1
2	1
5	1
4	1
3	0

Estimating Other Probabilities

- Recall the other probabilities:
  - **Marginals**: $\mathbb{P}(X = x)$ and $\mathbb{P}(Y = y)$.
  - **Conditionals**: $\mathbb{P}(X = x \mid Y = y)$ and $\mathbb{P}(Y = y \mid X = x)$.

- Can be calculated from the joint distribution.
  - Or an estimate of the joint distribution.

- Can also estimate more directly.

Estimating Marginals

- We estimate $\mathbb{P}(Y = y)$ with:

$$\mathbb{P}(Y = y) \approx \frac{\#(Y = y)}{n}$$

- E.g., estimate $\mathbb{P}(Y = 1)$:

- E.g., estimate $\mathbb{P}(Y = 0)$:

X	Y
5	0
3	0
4	1
4	1
2	0
5	1
2	1
5	1
4	1
3	0

Estimating Marginals

- We estimate $\mathbb{P}(X = x)$ with:

$$\mathbb{P}(X = x) \approx \frac{\#(X = x)}{n}$$

- E.g., estimate $\mathbb{P}(X = 4)$:

- E.g., estimate $\mathbb{P}(X = 3)$:

X	Y
5	0
3	0
4	1
4	1
2	0
5	1
2	1
5	1
4	1
3	0

Estimating Conditionals

- We estimate $\mathbb{P}(X = x \mid Y = y)$ with:

$$\mathbb{P}(X = x \mid Y = y) \approx \frac{\#(X = x \text{ and } Y = y)}{\#(Y = y)}$$

- E.g., estimate $\mathbb{P}(X = 4 \mid Y = 1)$:

- E.g., estimate $\mathbb{P}(X = 2 \mid Y = 0)$:

X	Y
5	0
3	0
4	1
4	1
2	0
5	1
2	1
5	1
4	1
3	0

Estimating Conditionals

- We estimate $\mathbb{P}(Y = y \mid X = x)$ with:

$$\mathbb{P}(Y = y \mid X = x) \approx \frac{\#(X = x \text{ and } Y = y)}{\#(X = x)}$$

- E.g., estimate $\mathbb{P}(Y = 1 \mid X = 4)$:
- E.g., estimate $\mathbb{P}(Y = 0 \mid X = 2)$:
- E.g., estimate $\mathbb{P}(Y = 0 \mid X = 6)$:

X	Y
5	0
3	0
4	1
4	1
2	0
5	1
2	1
5	1
4	1
3	0

Law of Large Numbers

- As data size $n \to \infty$, these esimated probabilities converge to their true expectations.[1]

[1] Assuming the data was sampled iid from the true distribution.

Bayes Classifier

- The Bayes classifier assumed we knew the true probabilities.
- But we can still use it if we replace the true probabilities with estimated probabilities.
- No longer guaranteed to be optimal!

Bayes Classifier

- Given a new flower with 5 petals, what is its class?
- Idea: estimate $\mathbb{P}(Y = 1 \mid X = 5)$.

X	Y
5	0
3	0
4	1
4	1
2	0
5	1
2	1
5	1
4	1
3	0

Multivariate Distributions

- We can also estimate when there are more variables in the same way.
- E.g., estimate $\mathbb{P}(Y = 1 \mid X_1 = 4, X_2 = 2)$:
- E.g., estimate $\mathbb{P}(X_1 = 2)$:
- E.g., estimate $\mathbb{P}(X_1 = 5, X_2 = 1 \mid Y = 1)$:

X_1	X_2	Y
5	1	0
3	3	0
4	2	1
4	5	1
2	3	0
5	2	1
2	1	1
5	1	1
4	2	0
3	6	0

DSC 140A

Probabilistic Modeling & Machine Learning

Lecture 12 | Part 2

Histogram Density Estimators

Continuous Variables

- We have seen how to estimate **discrete** probabilities. What about **continous** variables?
- Suppose there are two species of penguin; one species tends to have longer flippers.
- **Goal:** given a new penguin with flipper length $X = x$, predict its species, Y.

Data

- **Recall:** The distribution of a **continuous** random variable is described by a **density**.
- Can we estimate a density from data in the same way?
- E.g.: marginal density for x, $p_X(x)$. What is $p_X(7)$?

$$p_X(7) \overset{?}{\approx} \frac{\#(X = 7)}{n}$$

X	Y
7.2	0
11.3	1
8.0	1
5.1	0
5.6	1
12.3	1
13.1	1
10.9	0
12.0	1
5.0	0

Estimating Density

- Since X is continuous, most values of X are **never seen** in the data.
- We need to do some **smoothing**.
- One approach: **histogram estimators**.

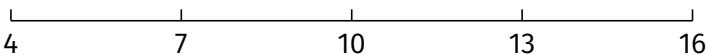
Histogram Estimators

- Suppose data $x_1, \ldots, x_n$ came from density f
- Divide domain into k **bins**: $[a_i, b_i)$.
  - Often equal-sized grid, though not necessary.
- Within each bin i, estimate density:

$$f(x) \text{ within bin } i \approx \frac{\#\text{ data points} \in [a_i, b_i)}{n \times \underbrace{(b_i - a_i)}_{\text{“bin width”}}}$$

Example

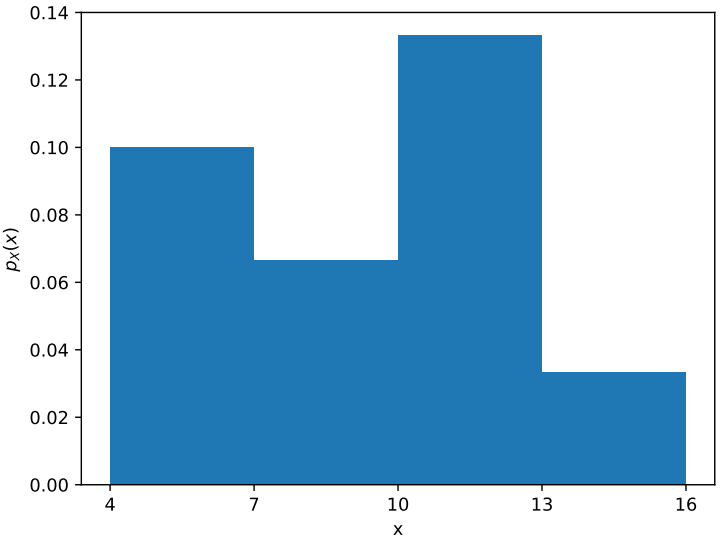
$$\frac{\#\text{ data points} \in [a_i, b_i)}{n \times (b_i - a_i)}$$



$[a_1, b_1) = [4, 7)$ $[a_2, b_2) = [7, 10)$ $[a_3, b_3) = [10, 13)$ $[a_4, b_4) = [13, 16)$

X	Y
7.2	0
11.3	1
8.0	1
5.1	0
5.6	1
12.3	1
13.1	1
10.9	0
12.0	1
5.0	0

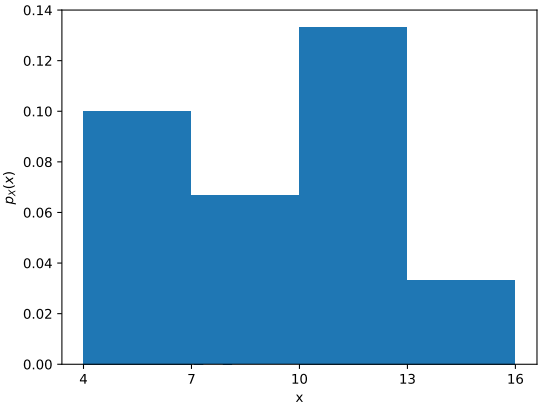
Histogram Estimator



X	Y
7.2	0
11.3	1
8.0	1
5.1	0
5.6	1
12.3	1
13.1	1
10.9	0
12.0	1
5.0	0

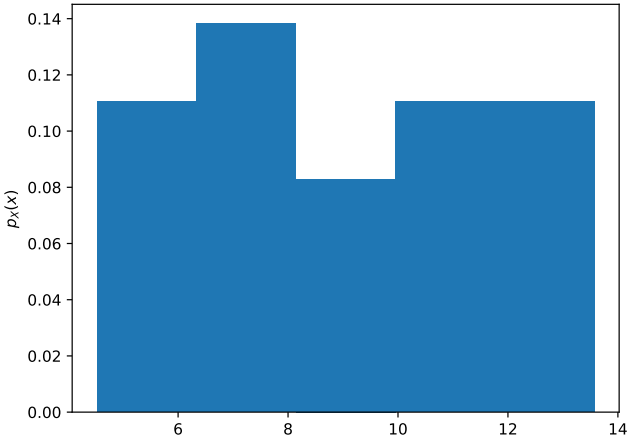
Histogram Estimator

- Histogram estimators produce density functions.
  - E.g., what is the estimated $p_X(4.7)$?
  - integrates (sums) to 1.



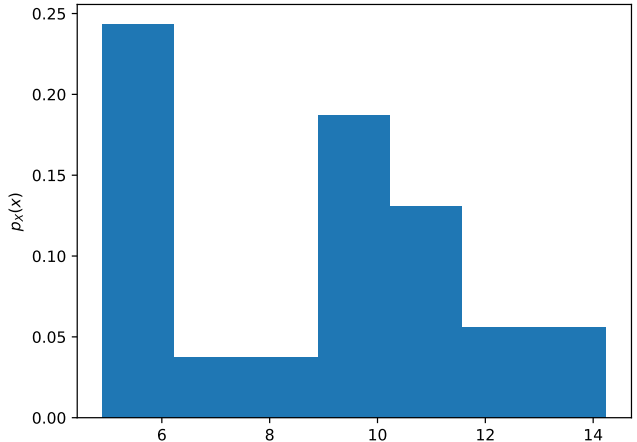
Bin Number and Sizes

- As we get more data, we can:
  - Decrease bin width.
  - Increase number of bins.



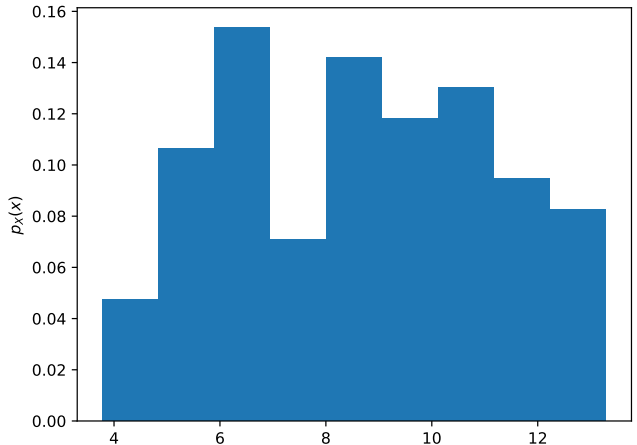
Bin Number and Sizes

- As we get more data, we can:
  - Decrease bin width.
  - Increase number of bins.



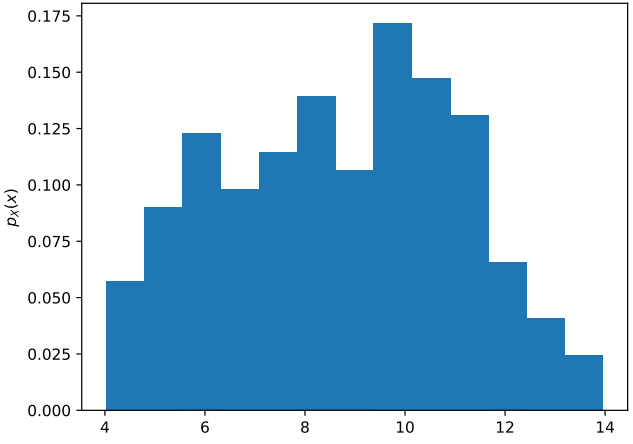
Bin Number and Sizes

- As we get more data, we can:
  - Decrease bin width.
  - Increase number of bins.



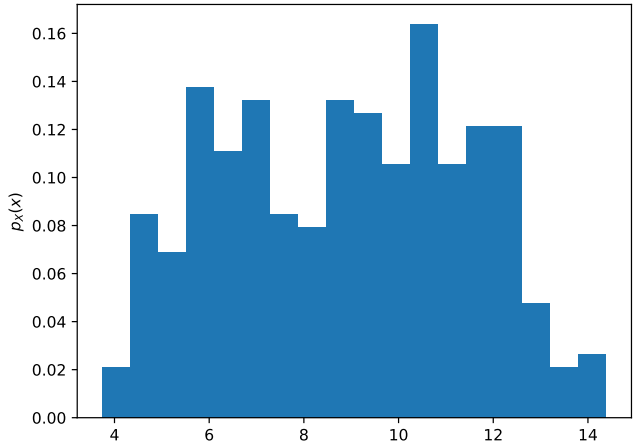
Bin Number and Sizes

- As we get more data, we can:
  - Decrease bin width.
  - Increase number of bins.



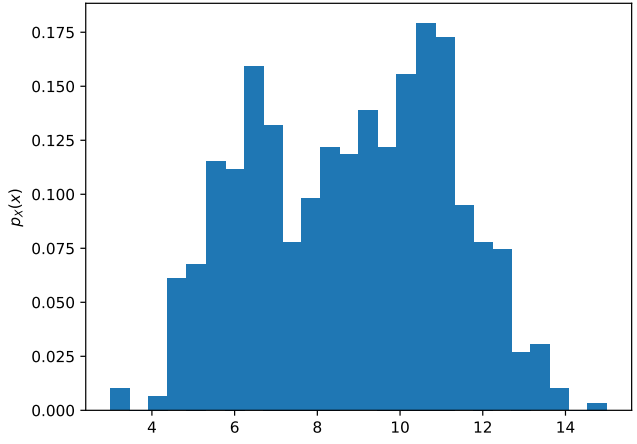
Bin Number and Sizes

- As we get more data, we can:
  - Decrease bin width.
  - Increase number of bins.



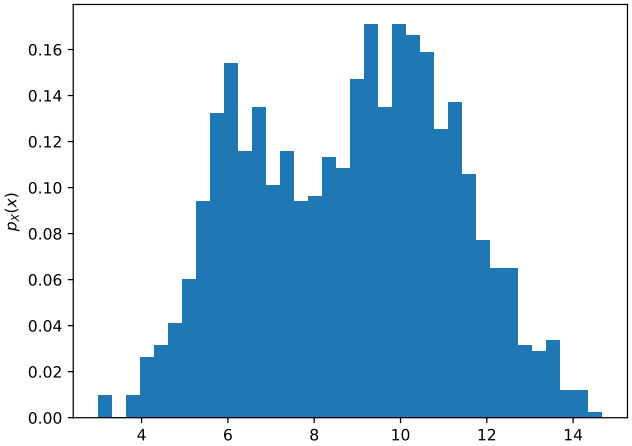
Bin Number and Sizes

- As we get more data, we can:
  - Decrease bin width.
  - Increase number of bins.



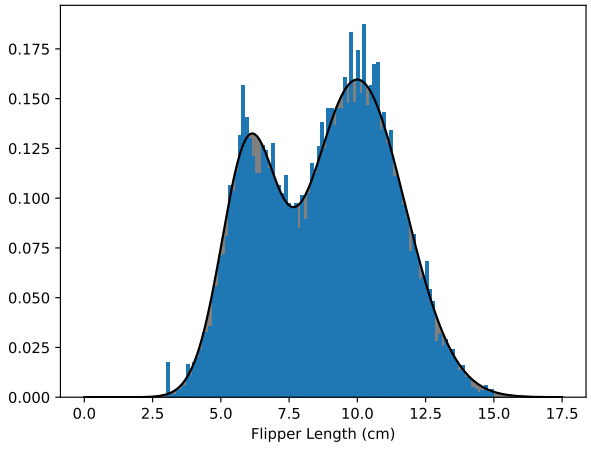
Bin Number and Sizes

- As we get more data, we can:
  - Decrease bin width.
  - Increase number of bins.



Law of Large Numbers

- Eventually, as n and # of bins $\to \infty$, the histogram estimator approaches the true density:

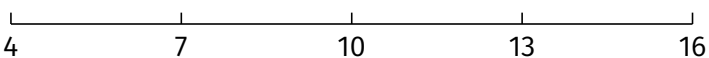


Estimating Conditional Distributions

- How do we estimate $p(x \mid Y = 1)$ and $p(x \mid Y = 0)$?
  - The flipper length densities for species 1 and 0.

- Restrict to data where $Y = 1$ (or $Y = 0$) and use histogram estimator.

Estimating $p(x \mid Y = y)$

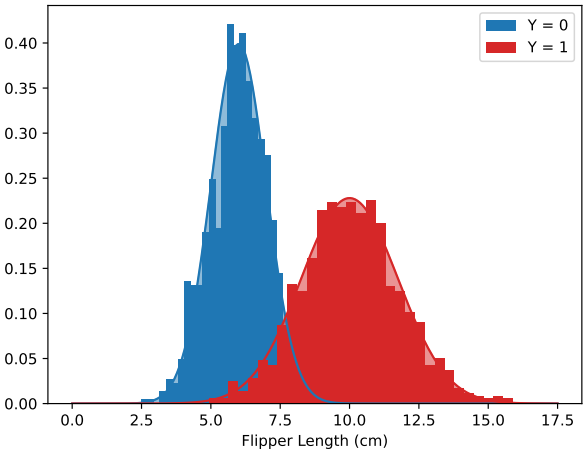


Estimate $p(x \mid Y = 0)$

X	Y
7.2	0
11.3	1
8.0	1
5.1	0
5.6	1
12.3	1
13.1	1
10.9	0
12.0	1
5.0	0

Law of Large Numbers

- Eventually, as n and # of bins $\to \infty$, the histogram estimators approach the true densities:



Estimating $\mathbb{P}(Y = y \mid X = x)$

- How do we estimate $\mathbb{P}(Y = y \mid X = x)$ with histograms?
- **Recall:** useful for making predictions.
- A discrete distribution, but conditioned on continuous variable.
  - Particular x may not be seen in data.

Estimating $\mathbb{P}(Y = y \mid X = x)$

- Two equivalent approaches:
  1. Count #($Y = 1$) and #($Y = 0$) within bin containing x.
  2. Compute from Bayes' rule and other estimates.

Approach #1: Directly

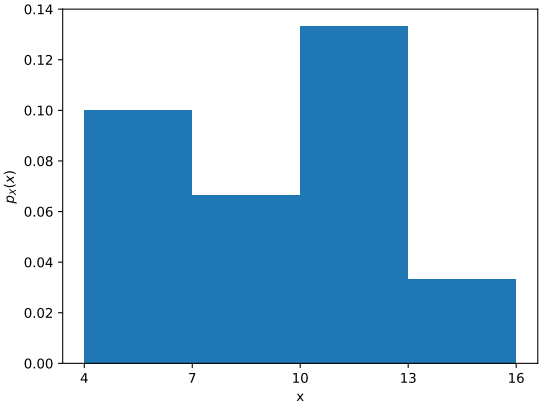
- To estimate $\mathbb{P}(Y = y \mid X = x)$ with histograms when Y is discrete and X is continuous:

1. Find the bin containing x.

2. Estimate:

$$\mathbb{P}(Y = y \mid X = x) \approx \frac{\#(Y = y \text{ within this bin })}{\#(\text{points within this bin})}$$

Approach #1: Directly



X	Y
7.2	0
11.3	1
8.0	1
5.1	0
5.6	1
12.3	1
13.1	1
10.9	0
12.0	1
5.0	0

Example: estimate $\mathbb{P}(Y = 1 \mid X = 4.3)$.

Approach #2: Bayes' Rule

1. Estimate other densities / probabilities:

$$p(x \mid Y = y) \qquad \mathbb{P}(Y = y) \qquad p_X(x)$$

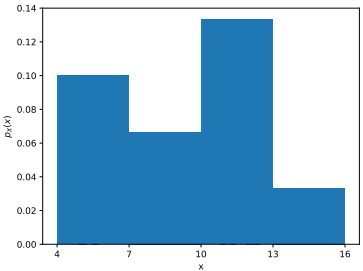
2. Use Bayes' rule to combine them:

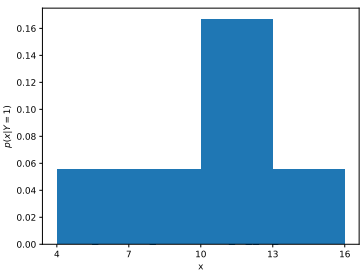
$$\mathbb{P}(Y = y \mid X = x) = \frac{p(x \mid Y = y)\mathbb{P}(Y = y)}{p_X(x)}$$

Approach #2: Bayes' Rule

- Using Bayes' rule:

$$\mathbb{P}(Y = y \mid X = x) = \frac{p(x \mid Y = y)\mathbb{P}(Y = y)}{p_X(x)}$$





Example: estimate $\mathbb{P}(Y = 1 \mid X = 4.3)$.

X	Y
7.2	0
11.3	1
8.0	1
5.1	0
5.6	1
12.3	1
13.1	1
10.9	0
12.0	1
5.0	0

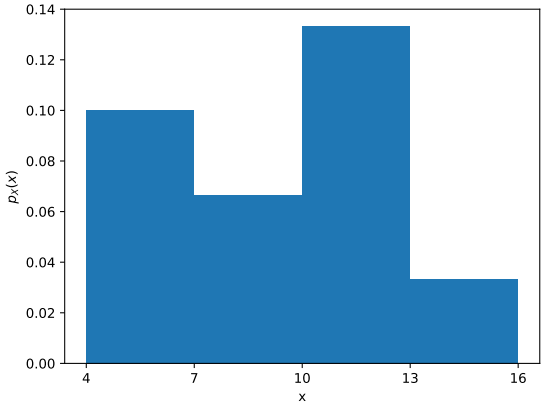
Equivalence

- Both approaches produce the same answer if same bins used to estimate all densities.
- Related via Bayes' rule.

Prediction

- Suppose there are two species of penguin; one species tends to have longer flippers.
- **Goal:** given a new penguin with flipper length $X = x$, predict its species, Y.

Example

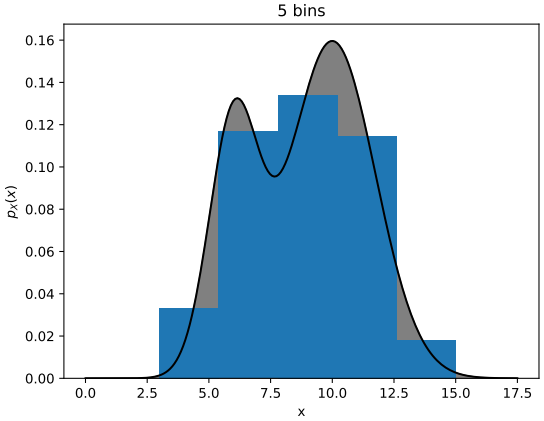


X	Y
7.2	0
11.3	1
8.0	1
5.1	0
5.6	1
12.3	1
13.1	1
10.9	0
12.0	1
5.0	0

Example: what is predicted species when X = 10.8?

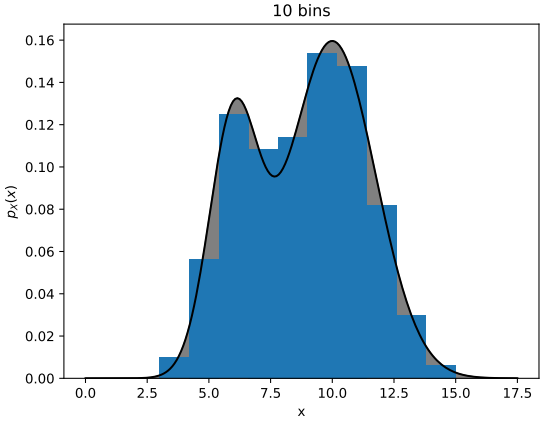
Over- and Under-fitting

- The number of bins must be chosen appropriately to avoid over- or under-fitting.



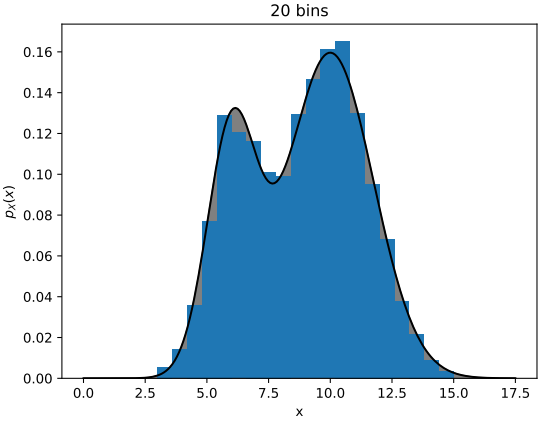
Over- and Under-fitting

- The number of bins must be chosen appropriately to avoid over- or under-fitting.



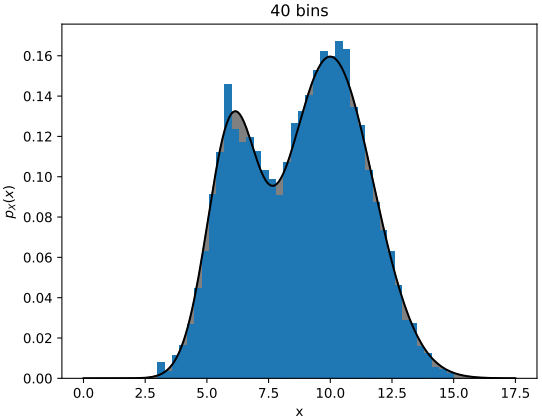
Over- and Under-fitting

- The number of bins must be chosen appropriately to avoid over- or under-fitting.



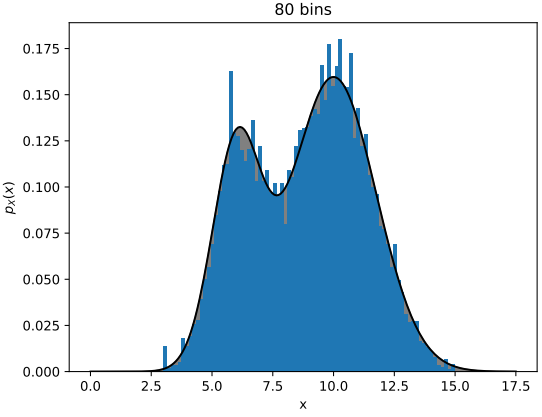
Over- and Under-fitting

- The number of bins must be chosen appropriately to avoid over- or under-fitting.



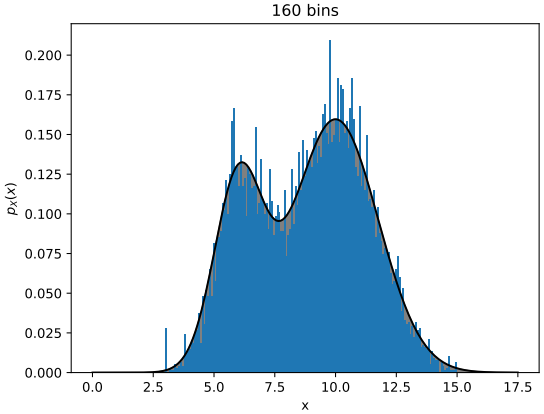
Over- and Under-fitting

- The number of bins must be chosen appropriately to avoid over- or under-fitting.



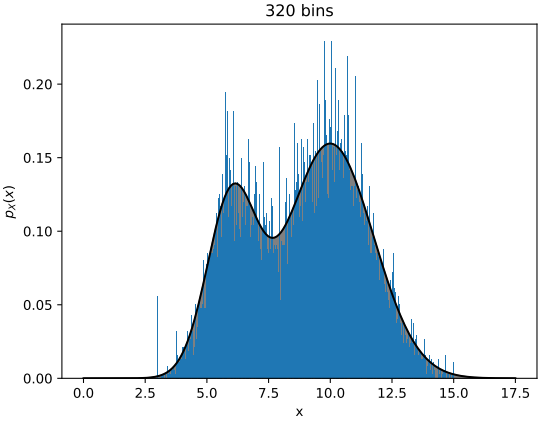
Over- and Under-fitting

- The number of bins must be chosen appropriately to avoid over- or under-fitting.



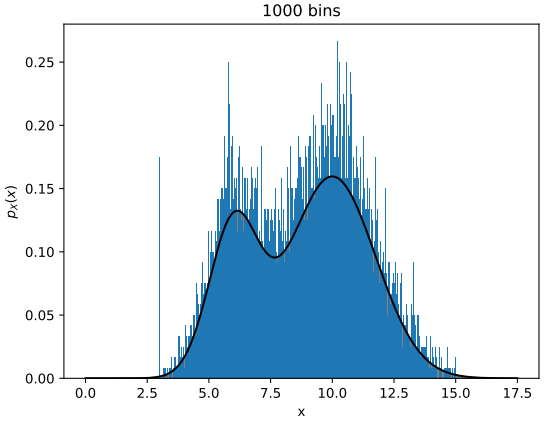
Over- and Under-fitting

- The number of bins must be chosen appropriately to avoid over- or under-fitting.



Over- and Under-fitting

- The number of bins must be chosen appropriately to avoid over- or under-fitting.



DSC 140A

Probabilistic Modeling & Machine Learning

Lecture 12 | Part 3

Multivariate Histogram Density Estimators

Multivariate Estimation

- In practice, we typically want to predict Y from many variables, X_1, X_2, ...
- How do we estimate densities $p(\vec{x})$ of several variables?

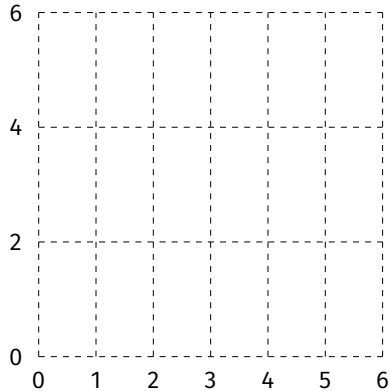
Histogram Estimators

- Histograms naturally generalize to $d > 1$:
- Suppose data $\vec{x}^{(1)}, \ldots, \vec{x}^{(n)}$ came from density f
- Divide $\mathbb{R}^d$ into rectangular bins **bins** with regular side-lengths $\ell_1, \ell_2, \ldots, \ell_d$
- Within a bin, estimate density:

$$f(\vec{x}) \text{ within bin} \approx \frac{\#\text{ data points} \in [a_i, b_i)}{n \times \underbrace{\left(\ell_1 \times \ell_2 \times \cdots \times \ell_d\right)}_{\text{“bin volume”}}}$$

Example: $d = 2$

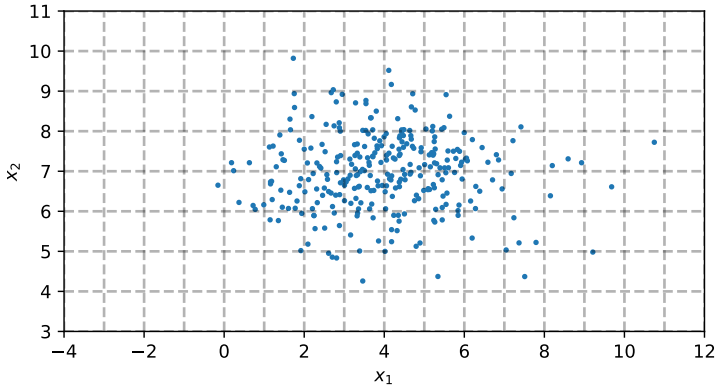
$$\frac{\#\text{ data points} \in [a_i, b_i)}{n \times (\ell_1 \times \ell_2 \times \cdots \times \ell_d)}$$



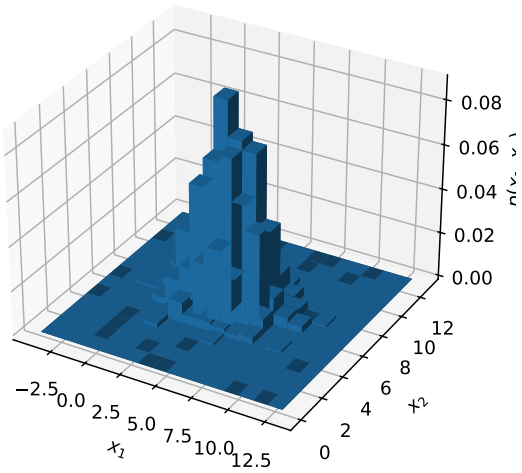
E.g., estimate: 1) $p_{X_1,X_2}(2.3, 2.5)$ 2) $p_{X_1,X_2}(3.3, 4.1)$

X_1	X_2	Y
4.1	1.8	0
3.6	3.0	0
4.2	2.2	1
4.2	2.4	1
2.3	3.2	0
4.9	2.4	1
2.1	0.8	1
3.2	1.1	1
4.7	2.3	0
3.8	4.9	0

Estimating 2-d Densities



Estimating 2-d Densities



Estimating in High Dimensions

- Histogram estimators can be used to estimate high-dimensional densities, *in principle*.
  - That is, densities of many continuous variables.

- But they typically do not work well due to the **curse of dimensionality**.

Curse of Dimensionality

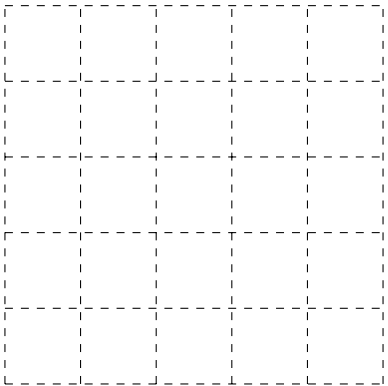
- Intuition: need sufficiently-many points in each bin to make good estimates.
  - Law of large numbers.
- Number of points needed is proportional to number of bins.
- **Many** bins in high dimensions.

Curse of Dimensionality

- Suppose we have two continuous variables, X_1 and X_2, each taking values between 0 and 1.
- Divide each feature into 5 equal bins:

0 0.2 0.4 0.6 0.8 1

Curse of Dimensionality



- Total number of bins: $5 \times 5 = 25$

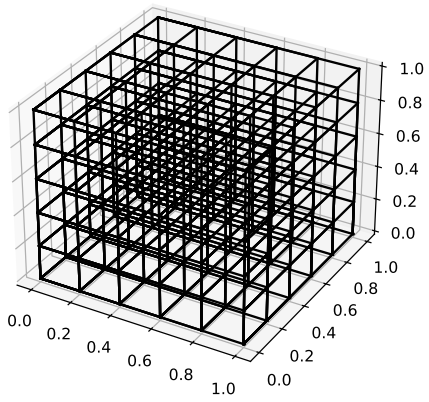
Curse of Dimensionality

- Suppose we have two continuous variables, X_1, X_2, X_3, each taking values between 0 and 1.

- Divide each feature into 5 equal bins:

0 0.2 0.4 0.6 0.8 1

Curse of Dimensionality



- Total number of bins: $5 \times 5 \times 5 = 5^3 = 125$

Curse of Dimensionality

- With d features, we'd have 5^d bins.
- Example: with 20 features, we'd have

$$5^{20} \approx 10 \text{ trillion}$$

Curse of Dimensionality

- To accurately estimate densities in more than a few dimensions, we need **too much data**.
- Most bins will be empty.
- And so we take different approaches.

A Different Approach

- Histogram estimators don't make assumptions about the **shape** of the density.
  - **Good**: very flexible.
  - **Bad**: requires a lot of data.

- **Next:** Assume a particular shape (e.g., a Gaussian) and try to learn it from data.