DSC 140A

Probabilistic Modeling & Machine Learning

Lecture 8 | Part 1

Estimating Discrete Probabilities

Recap: Bayes Classifier

- Bayes classifier: 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.

- **Optimal**: smallest possible probability of error.

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 $\mathbb{P}(X = 4, Y = 1)$?

$\approx 30\%$

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 $\mathbb{P}(X = 4, Y = 1)$: $\frac{3}{10} = .3$
- E.g., estimate $\mathbb{P}(X = 3, Y = 0)$: $\frac{2}{10} = .2$
- E.g., estimate $\mathbb{P}(X = 3, Y = 1)$: $\frac{0}{10} = .0$

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 $\mathbb{P}(Y = 1)$: $\frac{6}{10} = .6$

- E.g., estimate $\mathbb{P}(Y = 0)$: $\frac{4}{10} = .4$

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 $\mathbb{P}(X = 4)$: $\frac{3}{10} = .3$
- E.g., estimate $\mathbb{P}(X = 3)$: $\frac{2}{10} = .2$

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

$P(x|y) = P(x,y)/P(y)$

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 $\mathbb{P}(X = 4 \mid Y = 1)$: $\frac{3}{6} = .5$
- E.g., estimate $\mathbb{P}(X = 2 \mid Y = 0)$: $\frac{1}{4} = .25$

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 $\mathbb{P}(Y = 1 \mid X = 4)$: $\frac{3}{3} = 1.0$

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

0

- 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 values.[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 $\mathbb{P}(Y = 1 \mid X = 5)$. $= \frac{2}{3}$

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 $\mathbb{P}(Y = 1 \mid X_1 = 4, X_2 = 2)$: $\frac{1}{2}$
- E.g., estimate $\mathbb{P}(X_1 = 2)$: $\frac{2}{10}$
- E.g., estimate $\mathbb{P}(X_1 = 5, X_2 = 1 \mid Y = 1)$: $\frac{1}{5}$

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 8 | Part 2

Histogram Density Estimators

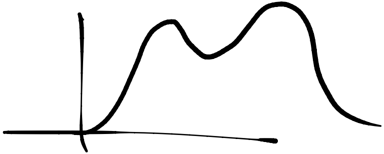
Continuous Variables

- 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

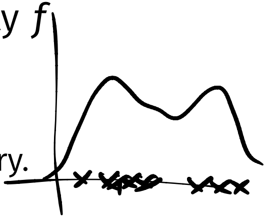
- Since X is continuous, most values of X are **never seen** in the data.

- We need to do some **smoothing**.

- One approach: **histogram estimators**.

Histogram Estimators

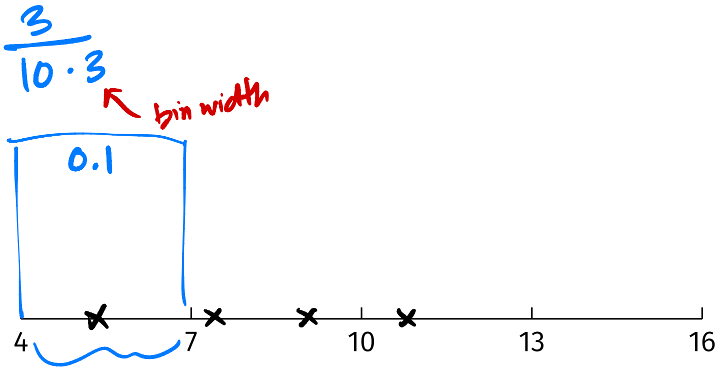
- Suppose data $x_1, \dots, x_n$ came from density f
- Divide domain into k **bins**: $[a_i, b_i)$.
  - Often equal-sized grid, though not necessary.



- Within each bin i, estimate density:

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

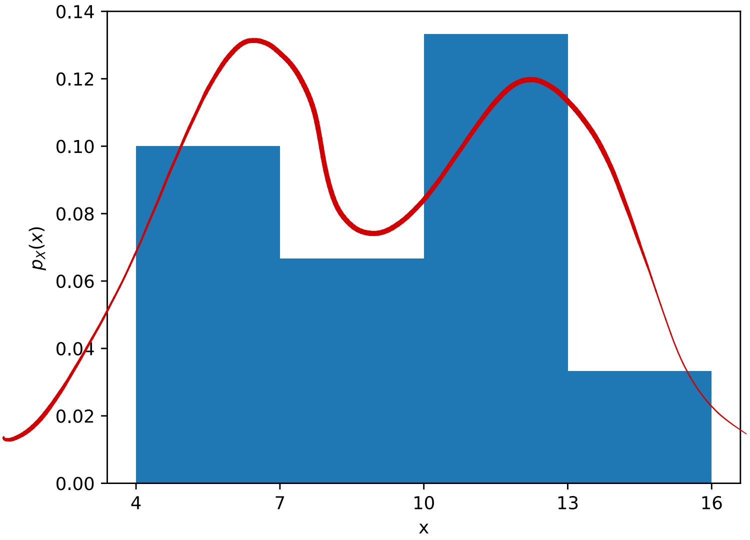
Example



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

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

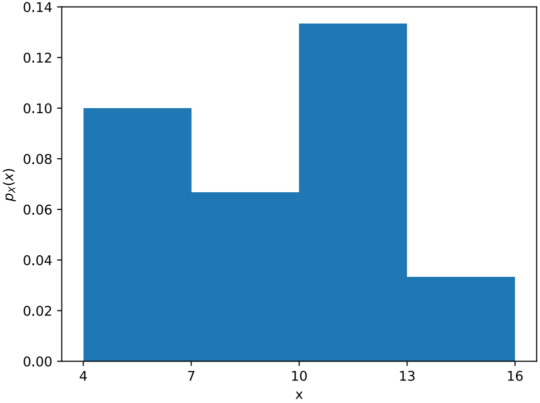
Histogram Estimator



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

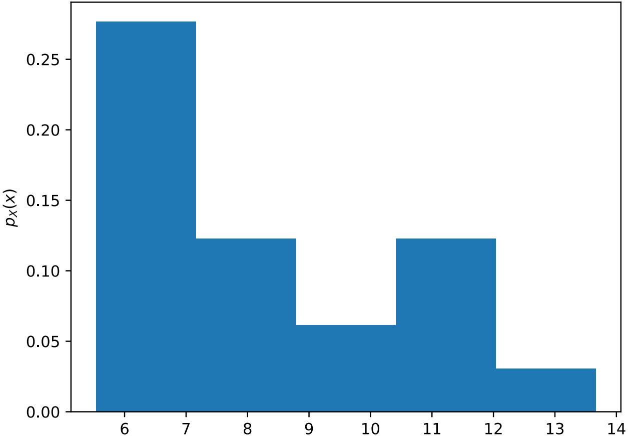
Histogram Estimator

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



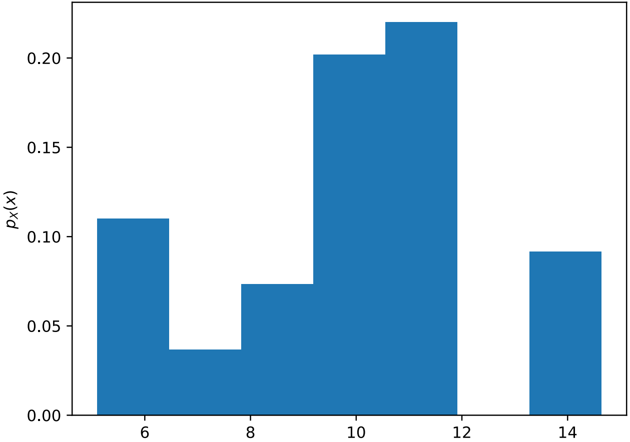
Bin Number and Sizes

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



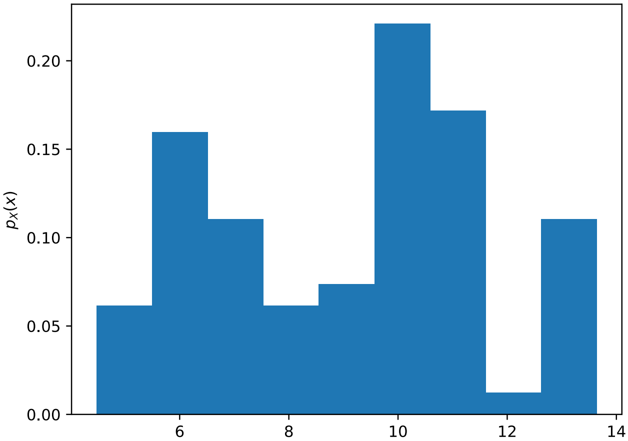
Bin Number and Sizes

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



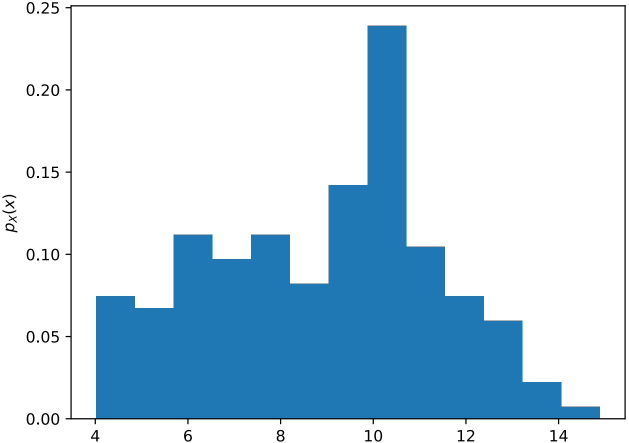
Bin Number and Sizes

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



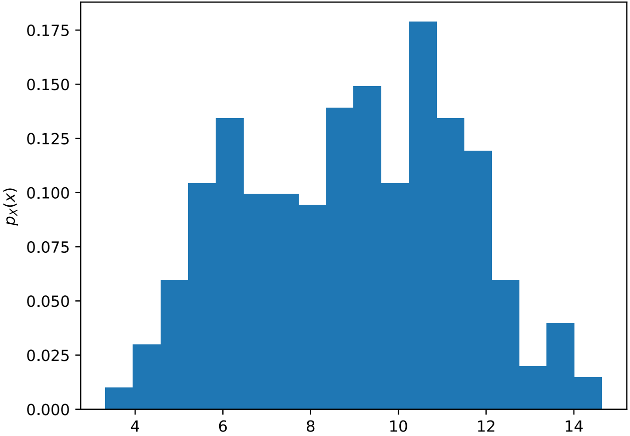
Bin Number and Sizes

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



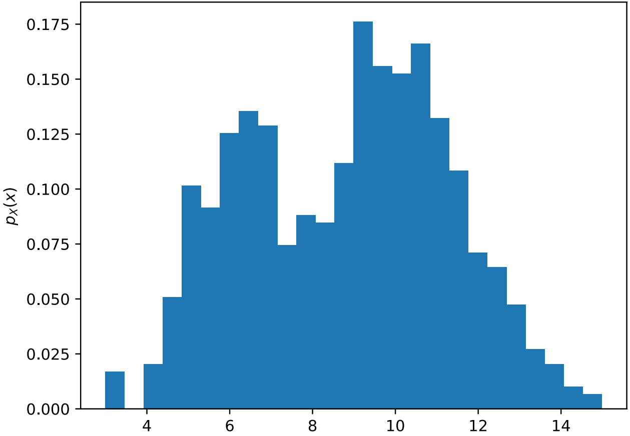
Bin Number and Sizes

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



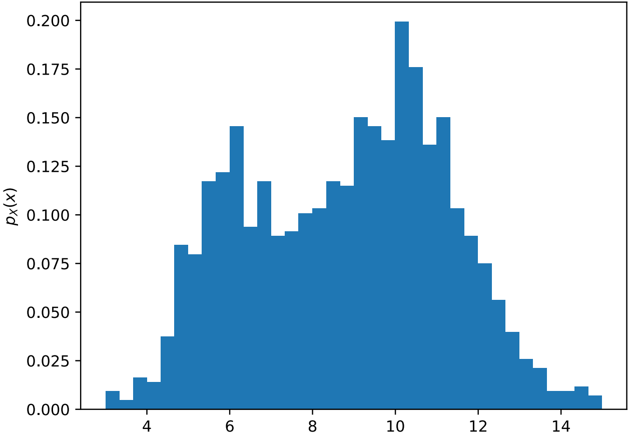
Bin Number and Sizes

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



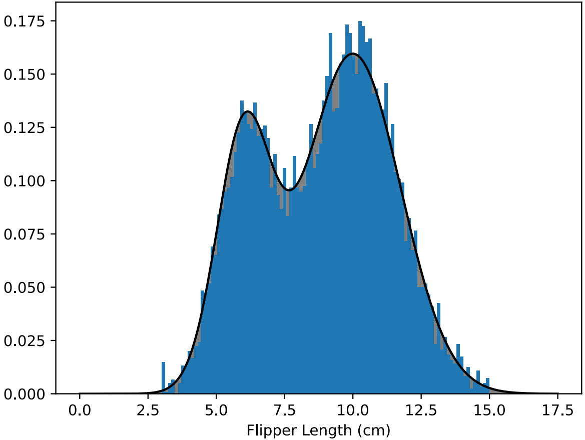
Bin Number and Sizes

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



Law of Large Numbers

- Eventually, as *n* and # of bins → ∞, the histogram estimator approaches the true density:

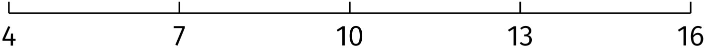


Estimating Conditional Distributions

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

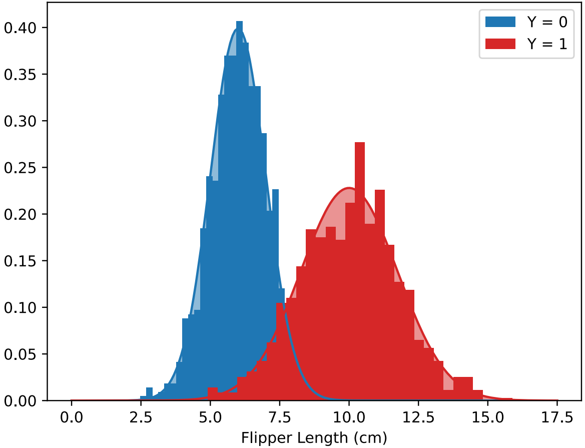


Estimate $p(x \mid Y = 0)$

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 → ∞, 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.

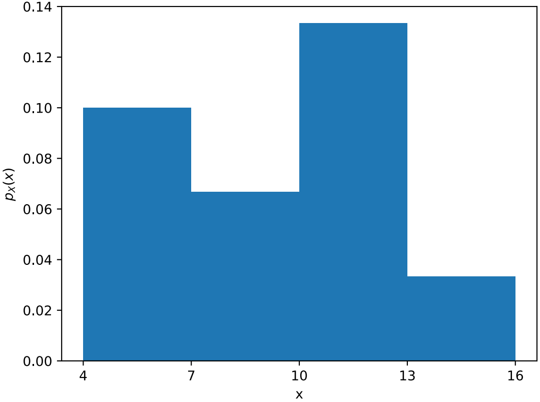
Approach #1: Directly

- 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 $\mathbb{P}(Y = 1 \mid X = 4.3)$. $\frac{1}{3}$

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:

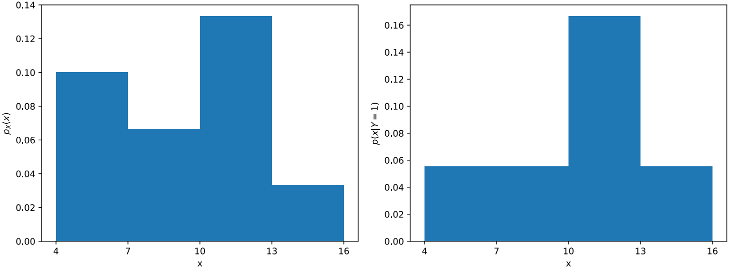
$$\mathbb{P}(Y = y \mid X = x) = \frac{p(x \mid Y = y)\mathbb{P}(Y = y)}{p_X(x)}$$

Approach #2: Bayes' Rule

$$\frac{1}{\cancel{6}\cdot 3}\cdot\frac{\cancel{6}}{\cancel{10}}\times\cancel{10}$$

- Using Bayes' rule:

$$\mathbb{P}(Y=\overset{1}{y}\mid X=x)=\frac{\overset{\frac{1}{6\times3}\times\frac{6}{10}}{p(x\mid Y=y)\,\mathbb{P}(Y=y)}}{\underset{\frac{1}{10}}{p_X(x)}}$$



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

Equivalence

- Both approaches produce the same answer if same bins used to estimate all densities.

- Related via Bayes’ rule.

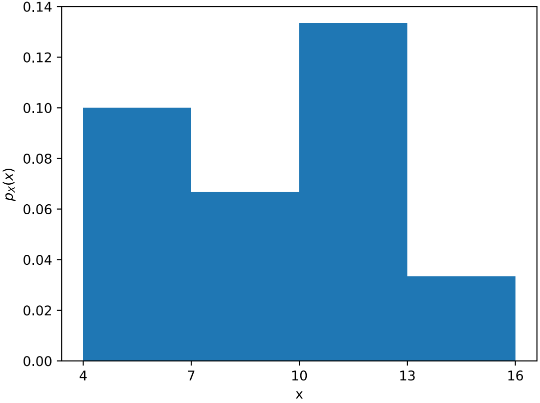
Prediction

- 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

$\mathbb{P}(Y=1 \mid X=10.8)$
$= 3/4$

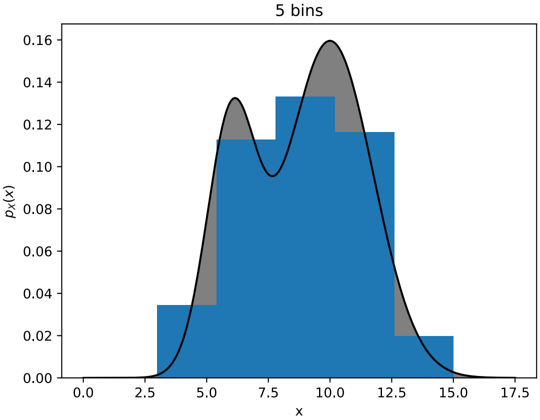


X	Y
7.2	0
11.3	1
8.0	1
5.1	0
5.6	1
12.3	1
13.1	1
10.9	0
12.0	1
5.0	0

Example: what is predicted species when X = 10.8?

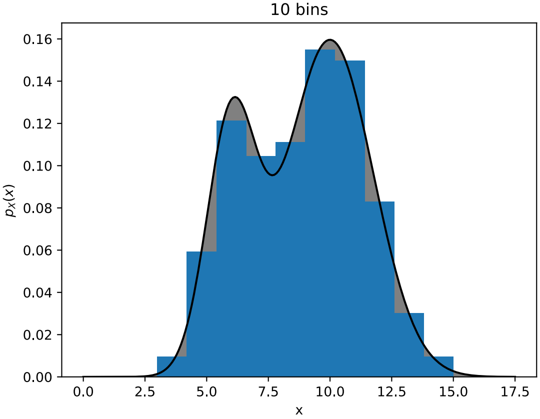
Over- and Under-fitting

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



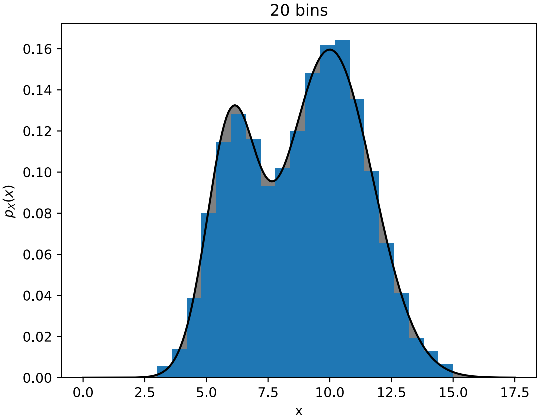
Over- and Under-fitting

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



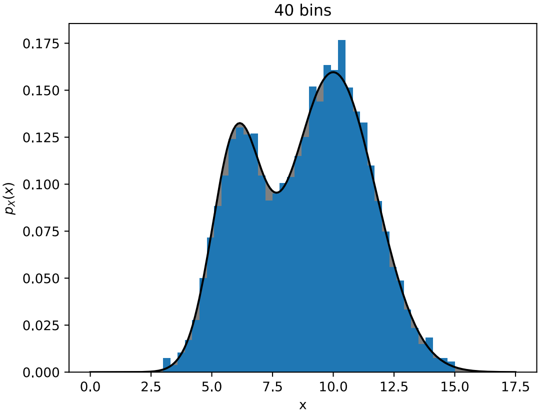
Over- and Under-fitting

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



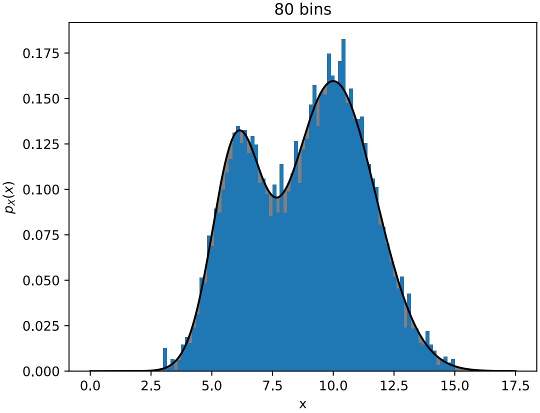
Over- and Under-fitting

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



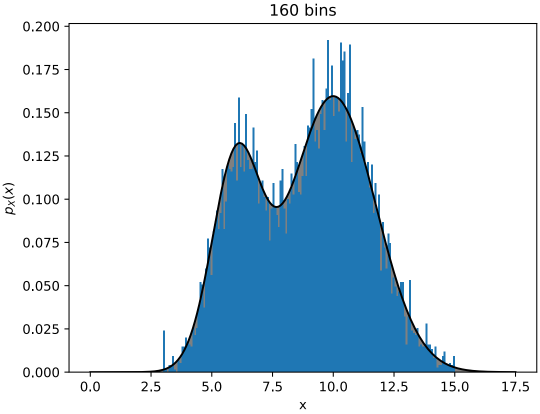
Over- and Under-fitting

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



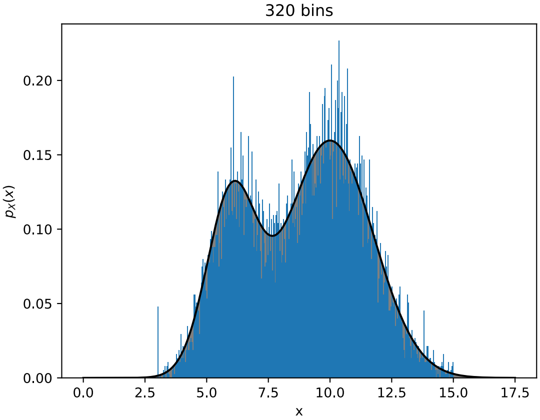
Over- and Under-fitting

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



Over- and Under-fitting

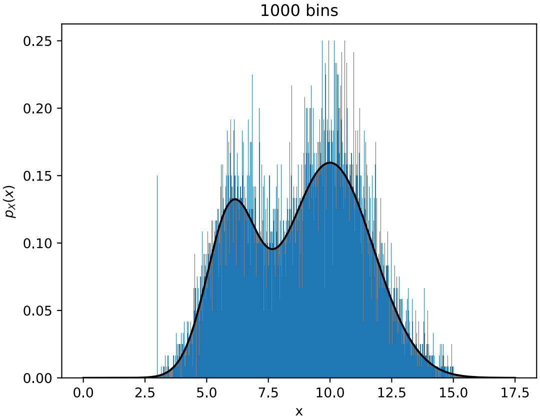
- The number of bins must be chosen appropriately to avoid over- or under-fitting.



Over- and Under-fitting

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

$\sqrt{n}$



DSC 140A

Probabilistic Modeling & Machine Learning

Lecture 8 | 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 $d > 1$:

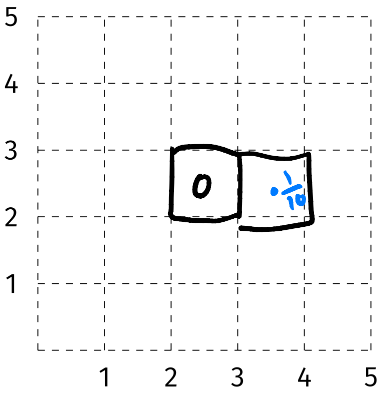
- Suppose data $\vec{x}^{(1)}, \ldots, \vec{x}^{(n)}$ came from density f

- 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

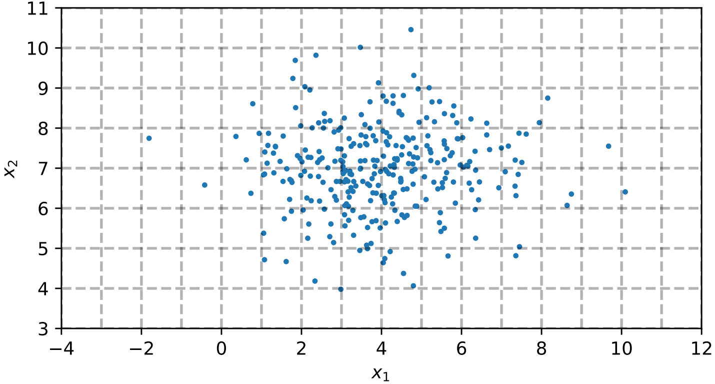


$\frac{1}{10(17}$

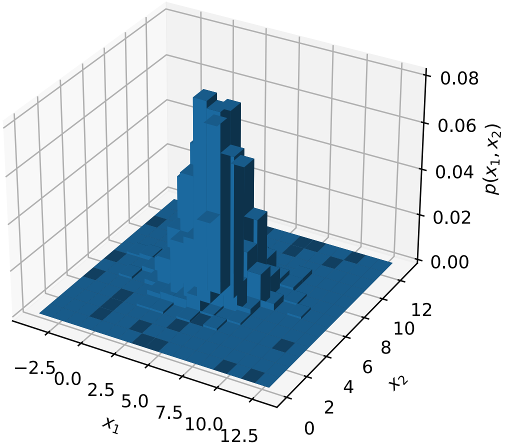
X_1	X_2	Y
4.1	1.8	0
→ 3.6	~~3.0~~ 2.9	0
4.2	2.2	1
4.2	2.4	1
2.3	3.2	0
4.9	2.4	1
2.1	0.8	1
3.2	1.1	1
4.7	2.3	0
3.8	4.9	0

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

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.

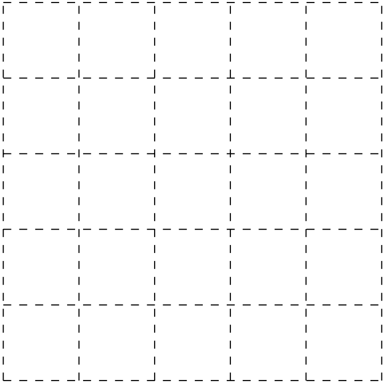
Curse of Dimensionality

- 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 × 5 = 25

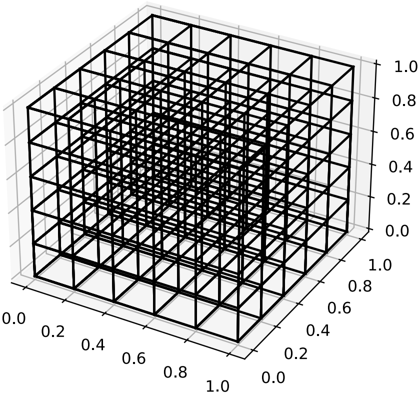
Curse of Dimensionality

- 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



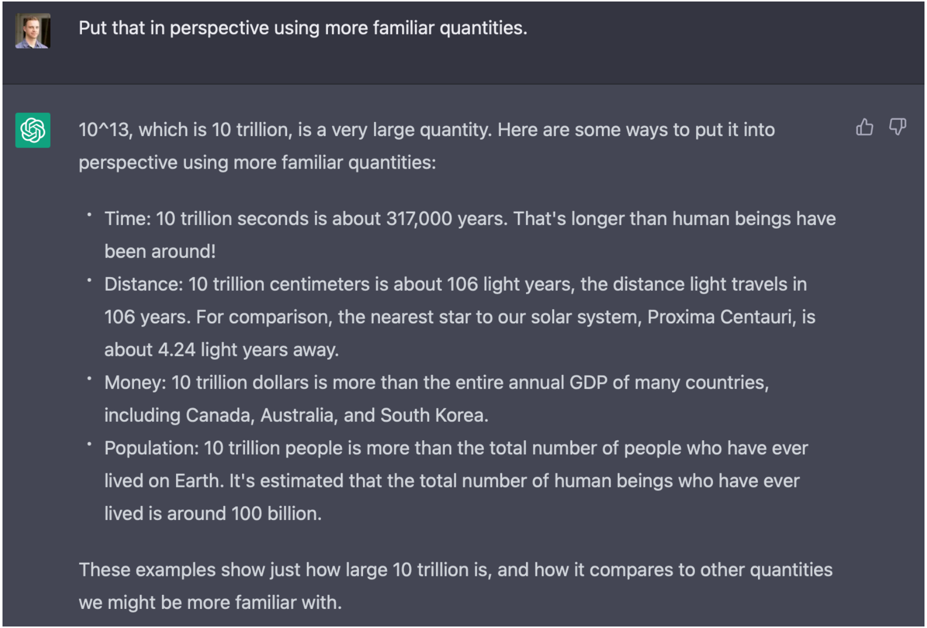
- Total number of bins: $5 \times 5 \times 5 = 5^3 = 125$

Curse of Dimensionality

- With d features, we'd have 5^d bins.

- Example: with 20 features, we'd have

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



Put that in perspective using more familiar quantities.

10^13, which is 10 trillion, is a very large quantity. Here are some ways to put it into perspective using more familiar quantities:

- Time: 10 trillion seconds is about 317,000 years. That's longer than human beings have been around!
- Distance: 10 trillion centimeters is about 106 light years, the distance light travels in 106 years. For comparison, the nearest star to our solar system, Proxima Centauri, is about 4.24 light years away.
- Money: 10 trillion dollars is more than the entire annual GDP of many countries, including Canada, Australia, and South Korea.
- Population: 10 trillion people is more than the total number of people who have ever lived on Earth. It's estimated that the total number of human beings who have ever lived is around 100 billion.

These examples show just how large 10 trillion is, and how it compares to other quantities we might be more familiar with.

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.