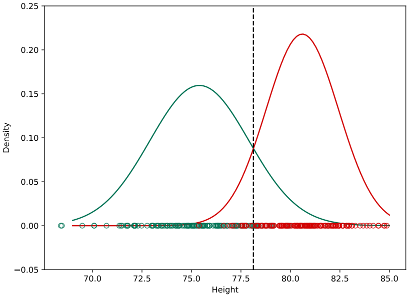


CSE 151A

Intro to Machine Learning

Lecture 03 – Part 01

The Probabilistic View

Recap

- Some tasks are not easily dictated to computers.
- Instead, we give it data and let it learn.
- But we still have to tell it how to learn.
- The magic is in the **data**.

Recap: Classification

- Predict which class the input belongs to.
- We encode instance as a **feature vector**.
- We have a **training set** of feature vectors and associated **class labels**.

Recap: Classification

- Train classifier to do well on the training data.
- Really want it to do well on data **we haven't seen**.
- That is, we want the classifier to **generalize**.

Recap: Classification

- We estimate ability to generalize with a **test set**.
- Test set contains the "right answers", too.
- Training error = error on training set.
- Test error = error on test set.

Recap: k-Nearest Neighbors

- kNN: “memorize” the training set.
- Predict the majority class of the k nearest neighbors in the training set.
- **Works well**, **simple**, **slow**, **memory-intensive**, **struggles in high dimensions**.

A New View via Probability

Example

- The average height of a forward is 80.5 inches.
- The average height of a guard is 75.4 inches.
- A new player is 73 inches tall. They're probably a guard / forward.

Example

- The average height of a forward is 80.5 inches.
- The average height of a guard is 75.4 inches.
- A new player is 73 inches tall. They're probably a **guard** / forward.

Example

- The average height of a forward is 80.5 inches.
- The average height of a guard is 75.4 inches.
- Another new player is 81 inches tall. They're probably a guard / forward.

Example

- The average height of a forward is 80.5 inches.
- The average height of a guard is 75.4 inches.
- Another new player is 81 inches tall. They’re probably a guard / **forward**.

Example

- The average height of a forward is 80.5 inches.
- The average height of a guard is 75.4 inches.
- Another new player is 77.5 inches tall. They're probably a guard / forward.

Example

- The average height of a forward is 80.5 inches.
- The average height of a guard is 75.4 inches.
- Another new player is 77.5 inches tall. They're probably a ???

Example

- Each new player has some **probability** of being a **guard** and some **probability** of being a **forward**.
- These probabilities are influenced by the player's **height**.

Conditional Probabilities

- Let X be player's height. Let Y be position.
- The probability that a player is a forward **given** they are 77.5 inches tall is written:

$$P(Y = \text{forward} | X = 77.5)$$

- This is the **conditional probability** of position given height.

Example

- $P(Y = \textbf{forward} | X = 81) \approx 1$
- $P(Y = \textbf{forward} | X = 70) \approx 0$
- $P(Y = \textbf{forward} | X = 77.5) \approx \frac{1}{2}$?
- $P(Y = \textbf{guard} | X = 81) \approx 0$
- $P(Y = \textbf{guard} | X = 70) \approx 1$
- $P(Y = \textbf{guard} | X = 77.5) \approx \frac{1}{2}$?

Classification

- A new player is x inches tall. What is their position?
- If $P(Y = \text{forward}|X = x) > P(Y = \text{guard}|X = x)$, predict **forward**.
- If $P(Y = \text{guard}|X = x) > P(Y = \text{forward}|X = x)$, predict **guard**.

The Bayes Classifier

- Given $X = x$, and possible classes $y_1, \dots, y_k$...
- ...predict the class y_i that makes $P(Y = y_i | X = x)$ the largest.

Bayes Error

- Assume new player with height x is either a guard or a forward (binary classification).
- Suppose $P(\text{guard}|X = x) = 0.6$.
- Then $P(\text{forward}|X = x) = 0.4$.
- We predict **guard**, but 40% chance we're wrong.

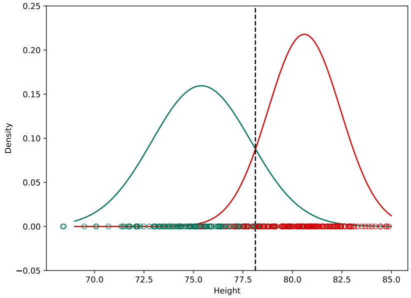
Bayes Error

- Usually, some error is unavoidable.
- Out of all possible classifiers, the Bayes Classifier has the smallest expected error.
- This error is called the **Bayes Error**.

Great! So ML is solved...

Problem: we don’t **know** these probabilities.

Solution: gather **data** and **estimate** them.



CSE 151A
Intro to Machine Learning

Lecture 03 – Part 02

Estimating Probabilities

Estimating Probabilities

- What is the probability that a tablet is defective?
- There is a “true” probability p; $P(\text{defective}) = p$.
- Estimate: sample n tablets, count # defective.

$$P(\text{defective}) \approx \frac{\#\,\text{defective}}{n}$$

Estimating Probabilities

- What is the probability that a tablet is defective?
- There is a "true" probability p; $P(\text{defective}) = p$.
- Estimate: sample n tablets, count # defective.

$$P(\text{defective}) \approx \frac{\#\,\text{defective}}{n}$$

- Law of large numbers: estimate $\to p$ as $n \to \infty$.

Estimating Conditional Probabilities

- What is the probability that a tablet is defective **given** that it is made by **Apple**?

- Use above data but discard non-Apple tablets:

$$P(\text{defective} \mid \text{Apple}) \approx \frac{\#\ \text{defective Apple}}{\#\ \text{Apple}}$$

Estimating Conditional Probabilities

- We estimate $P(A = a \mid B = b)$ by gathering data and counting:

$$\frac{\#\text{ for which } A = a \text{ and } B = b}{\#\text{ for which } B = b}$$

- **Problem**: what if A or B are **continuous**?

Example: Discrete A, Continuous B

- Estimate probability that player is a forward, given their height is 77.15 inches.

$$\frac{\text{\# forwards with height} = 77.15}{\text{\# height} = 77.15}$$

- **Problem**: no one in data w/ height *exactly* 77.15 in
- Divide by zero. **Undefined**.

Example: Continuous A, Discrete B

- Estimate probability that height = 77.15 in, given that player is a forward.

$$\frac{\text{\# forwards with height 77.15 in}}{\text{\# forwards}}$$

- **Problem**: no one in data w/ height *exactly* 77.15 in

- **Zero** unless 77.15 in. forward is in data set.

Solution: Smoothing

- If A is continuous, estimate

$$P(A = \text{close to } a \mid B = b)$$

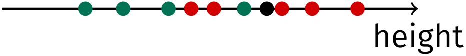
- If B is continuous, estimate

$$P(A = a \mid B = \text{close to } b)$$

- We'll see approaches for each case.

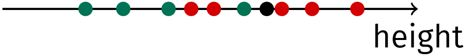
Discrete A, Continuous B

- Estimate $P(Y = \text{forward} \mid X = 77.15)$



Discrete A, Continuous B

- Estimate $P(Y = \text{forward} \mid X = 77.15)$

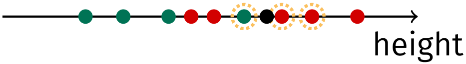


- Use fraction of $k = 3$ nearest neighbors:

$$P(Y = \text{forward} \mid X = 77.15) \approx \frac{\textbf{2 red}}{3} = \frac{2}{3}$$

Discrete A, Continuous B

- Estimate $P(Y = \text{forward} \mid X = 77.15)$

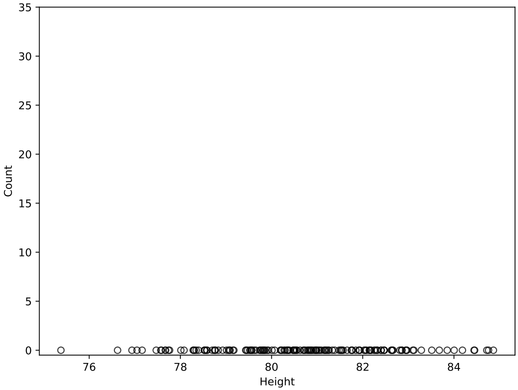


- Use fraction of $k = 3$ nearest neighbors:

$$P(Y = \text{forward} \mid X = 77.15) \approx \frac{\textbf{2 red}}{3} = \frac{2}{3}$$

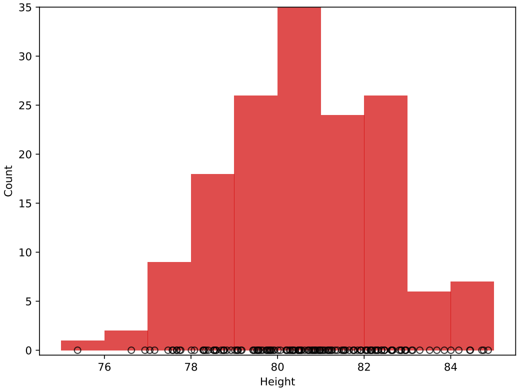
Continuous A, Discrete B

- Estimate $P(X = 77.15 \mid Y = \text{forward})$



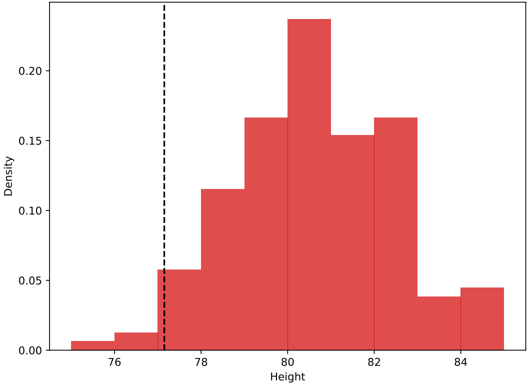
Continuous A, Discrete B

- Estimate $P(X = 77.15 \mid Y = \text{forward})$



Continuous A, Discrete B

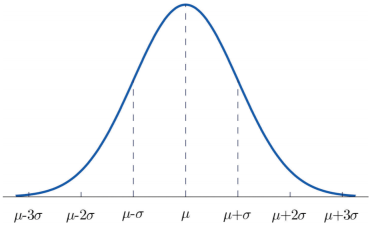
- Estimate $P(X = 77.15 \mid Y = \text{forward})$



Histograms

- We can estimate these probabilities with a **histogram**.
- Observe: the histogram is bell shaped.
- Let's try fitting a Normal curve.

The Normal Curve

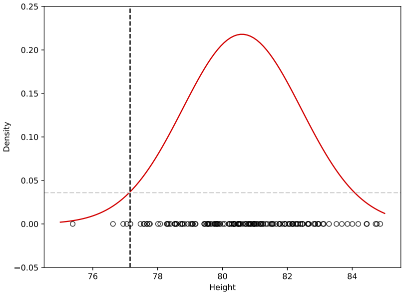


- The equation of the **univariate Gaussian**:

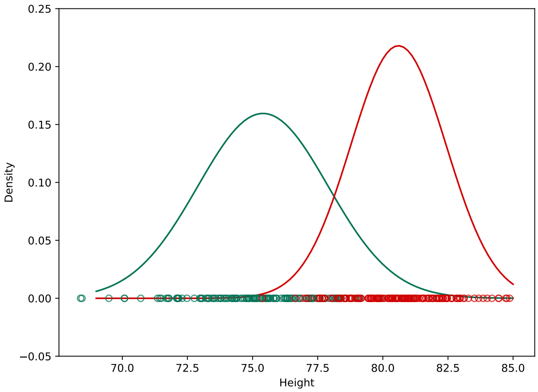
$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

Fitting a Normal Curve

- Calculate mean μ and STD σ from data.
- For forwards: $\mu = 80.5$, $\sigma = 1.84$



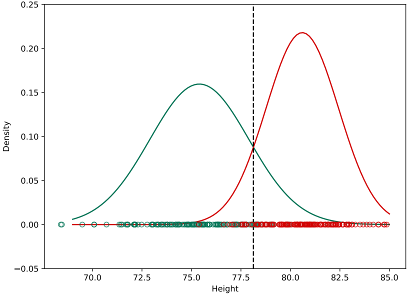
Fitting a Normal Curve to Each Class



Continuous A, Discrete B

- Two approaches: histograms and fitting a Gaussian.
- Histograms are **non-parametric**. No assumptions on shape.
- Fitting a Gaussian is **parametric**. Strong assumption.
- Other approaches exist.

Up next: using these estimates in the Bayes classifier.



CSE 151A

Intro to Machine Learning

Lecture 03 – Part 03

Using the Bayes Classifier

Remember: The Bayes Classifier

- Given $X = x$, and possible classes $y_1, \dots, y_k$...
- ...predict the class y_i that makes $P(Y = y_i | X = x)$ the largest.

Bayes Classifier and Estimation

- The Bayes Classifier is optimal if **true** probabilities are used.
- If **estimated** probabilities are substituted, the classifier is no longer **optimal**, but still **good**.

Roadmap

We'll see two approaches:

1. Estimating $P(Y|X)$.
2. Estimating $P(X|Y)$ & $P(Y)$ and using Bayes' Rule.

Approach #1: Estimating P(Y|X)

- A new player's height is 77.15 inẆhat is their position?

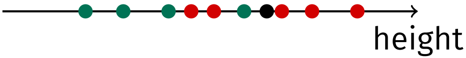
- We need to estimate

$$P(Y = \text{Forward} \mid X = 77.15)$$
$$P(Y = \text{Guard} \mid X = 77.15)$$

and choose the largest.

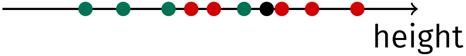
Discrete A, Continuous B

- Estimate $P(Y = \text{forward} \mid X = 77.15)$



Discrete A, Continuous B

- Estimate $P(Y = \text{forward} \mid X = 77.15)$

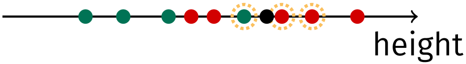


- Use fraction of $k = 3$ nearest neighbors:

$$P(Y = \text{forward} \mid X = 77.15) \approx \frac{\textbf{2 red}}{3} = \frac{2}{3}$$

Discrete A, Continuous B

- Estimate $P(Y = \text{forward} \mid X = 77.15)$



- Use fraction of $k = 3$ nearest neighbors:

$$P(Y = \text{forward} \mid X = 77.15) \approx \frac{\textbf{2 red}}{3} = \frac{2}{3}$$

Does this seem familiar?

- We predict **forward** because the majority of the k neighbors are **red**.
- This is just the **k-Nearest Neighbor** classifier.

In fact...

- **Theorem**: the 1NN classifier has at most **twice** the Bayes Error as $n \to \infty$.

Approach #2: Use Bayes’ Theorem

- Remember Bayes’ Theorem:

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

Bayes Classifier after Bayes' Rule

- Before: predict y_i maximizing $P(Y = y_i \mid X = x)$.
- Bayes' Rule says that this is equivalent to picking y_i to maximize:

$$\frac{P(X = x \mid Y = y_i)P(Y = y_i)}{P(X = x)}$$

A Simplification...

- Only the numerator will change with different y_i.
- So pick y_i to maximize:

$$P(X = x \mid Y = y_i)P(Y = y_i)$$

- We know how to estimate both terms.

Example: Using Histograms

- A new player's height is 77.15 in. What is their position? (Assume binary classification.)

- We need to estimate

$$P(X = 77.15 \mid Y = \text{forward})$$
$$P(X = 77.15 \mid Y = \text{guard})$$
$$P(Y = \text{forward})$$
$$P(Y = \text{guard})$$

Example: Using Histograms

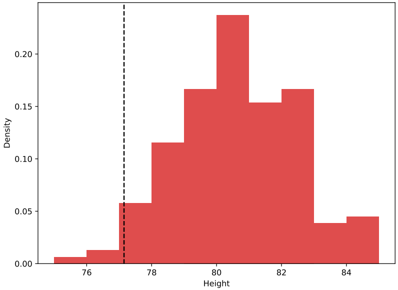
- Gather a training set of 300 players; 156 forwards, 144 guards.

$$P(Y = \text{forward}) \approx \frac{156}{300} = 0.52$$

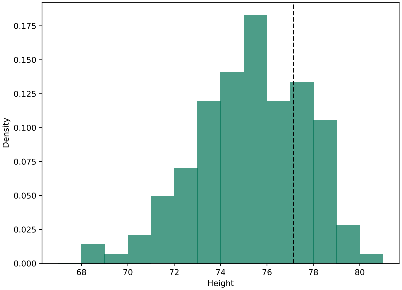
$$P(Y = \text{guard}) \approx \frac{144}{300} = 0.48$$

Example: Using Histograms

- Estimate conditional probs. with histograms.



$P(X = 77.15 \mid Y = \text{Forward}) \approx .057$



$P(X = 77.15 \mid Y = \text{Guard}) \approx .134$

Example: Using Histograms

- Therefore:

$$P(X = 77.15 \mid Y = \text{forward})P(Y = \text{forward}) \approx (.057)(.52) = .03$$
$$P(X = 77.15 \mid Y = \text{guard})P(Y = \text{guard}) \approx (.134)(.48) = .07$$

- Therefore, we predict **guard**.

- **Note**: probabilities do not add to one.

Example: Using Gaussians

- A new player's height is 77.15 in. What is their position? (Assume binary classification.)

- We need to estimate

$$P(X = 77.15 \mid Y = \text{forward})$$
$$P(X = 77.15 \mid Y = \text{guard})$$
$$P(Y = \text{forward})$$
$$P(Y = \text{guard})$$

Example: Using Gaussians

- Gather a training set of 300 players; 156 forwards, 144 guards.

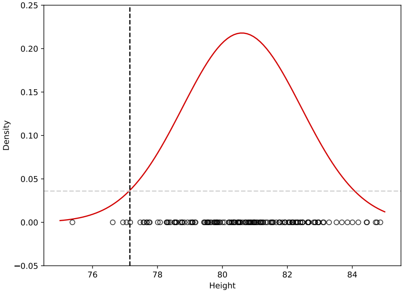
$$P(Y = \text{forward}) \approx \frac{156}{300} = 0.52$$

$$P(Y = \text{guard}) \approx \frac{144}{300} = 0.48$$

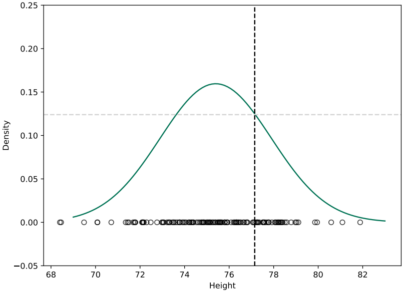
Example: Using Gaussians

- Estimate conditional probs. with Gaussians.
- For forwards: $\mu = 80.5, \sigma = 1.84$
- For guards $\mu = 75.4, \sigma = 2.5$
- Fit a Guassian for each.

Example: Using Gaussians



$P(X = 77.15 \mid Y = \text{Forward}) \approx .037$



$P(X = 77.15 \mid Y = \text{Guard}) \approx .125$

Example: Using Gaussians

- Therefore:

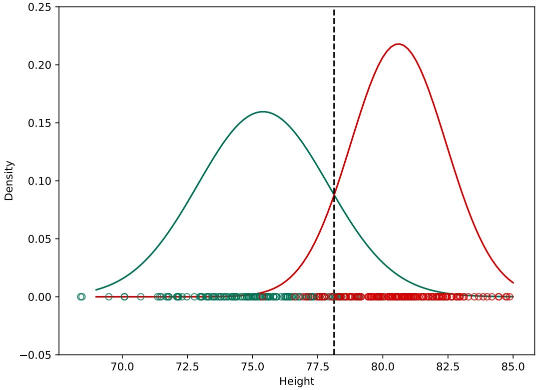
$$P(X = 77.15 \mid Y = \text{forward})P(Y = \text{forward}) \approx (.037)(.52) = .019$$
$$P(X = 77.15 \mid Y = \text{guard})P(Y = \text{guard}) \approx (.125)(.48) = .06$$

- Therefore, we predict **guard**.

- **Note**: probabilities do not add to one.

The Decision Boundary

- Plot $P(X = x \mid Y = \textbf{forward})P(Y = \textbf{forward})$ and $P(X = x \mid Y = \textbf{guard})P(Y = \textbf{guard})$



Results

- Train Error: 11.3%
- Test Error: 10%
- Best possible test error with simple boundary: 9.7%

Summary

- Estimate conditional probabilities with histogram estimators or by fitting Gaussians.
- Pick y_i to maximize:

$$P(X = x \mid Y = y_i)P(Y = y_i)$$

- This is called the **generative** approach to classification.

Next time...

- We have only used one feature so far (height).
- How do we include more?