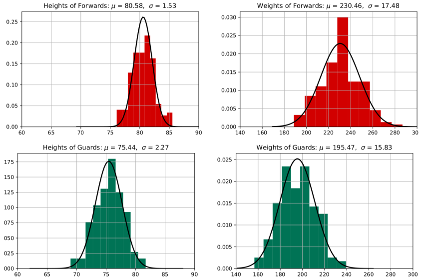


CSE 151A

Intro to Machine Learning

Lecture 05 – Part 01

What is Conditional Independence?

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

Remember: Independence

- Events A and B are **independent** if

$$P(A, B) = P(A) \cdot P(B).$$

- Equivalently, A and B are independent if[1]

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

[1] or $P(B) = 0$

Informally

- A and B are **independent** if learning B does not influence your belief that A happens.

Example

You draw one card from a deck of 52 cards. A is the event that the card is a heart, B is the event that the card is a face card (J,Q,K,A). Are these independent?

♥: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A
♦: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A
♣: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A
♠: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A

Example

We've lost the King of Clubs! You draw one card from this deck of 51 cards. A is the event that the card is a heart, B is the event that the card is a face card (J,Q,K,A). Are these independent?

♥: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A
♦: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A
♣: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, A
♠: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A

In the Real World...

- ...true independence is rare.
- Example, survivors of the titanic:

PassengerID	Survived	Pclass	Sex	Age	Fare	Embarked	FavColor
0	0	3	female	23.0	7.9250	S	yellow
1	0	1	male	47.0	52.0000	S	purple
2	0	3	male	36.0	7.4958	S	green
3	0	3	male	31.0	7.7500	Q	purple
4	0	3	male	19.0	7.8958	S	purple
...	...	...	...	...	...	...	...

In the Real World...

- $P(\text{Survived} = 1) = .408$
- $P(\text{Survived} = 1 \mid \text{FavColor} = \text{purple}) = .4$
- **Not independent...**

In the Real World...

- $P(\text{Survived} = 1) = .408$
- $P(\text{Survived} = 1 \mid \text{FavColor} = \text{purple}) = .4$
- **Not independent... ...but “close”!**

In the Real World...

- $P(\text{Survived} = 1) = .408$
- $P(\text{Survived} = 1 \mid \text{Pclass} = 1) =$

In the Real World...

- $P(\text{Survived} = 1) = .408$
- $P(\text{Survived} = 1 \mid \text{Pclass} = 1) = .657$

In the Real World...

- $P(\text{Survived} = 1) = .408$
- $P(\text{Survived} = 1 \mid \text{Pclass} = 1) = .657$
- **Strong dependence.**

Remember: Conditional Independence

- Events A and B are **conditionally independent** given C if

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

- Equivalently[2]:

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

[2] Or $P(B) = 0$

Informally

- Suppose you know that C has happened.
- You have some belief that A happens, given C.
- A and B are **conditionally independent** given C if learning that B happens in addition to C does not influence your belief that A happens given C.

Very informally

- A and B are **conditionally independent** given C if learning that B happens in addition to C gives you no more information about A.

Example

We've lost the King of Clubs! You draw one card from this deck of 51 cards. A is the event that the card is a heart, B is the event that the card is a face card (J,Q,K,A). Now suppose you know that the card is red. Are A and B independent **given** this information?

♥: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A
♦: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A
♣: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, A
♠: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A

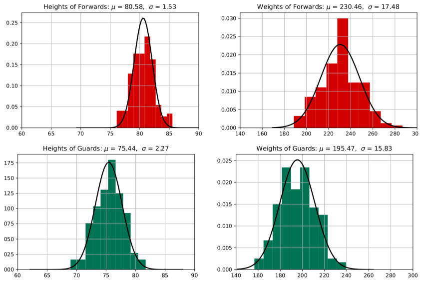
Titanic Example

- Survival and class are **not** independent.
- But they're (close) to **conditionally independent** given ticket price:
  - $P(\text{Survived} = 1 \mid \text{PClass} = 1, \text{Fare} > 50) = .708$
  - $P(\text{Survived} = 1 \mid \text{Fare} > 50) = .696$

More Variables

- $X_1, X_2, \ldots, X_d$ are **mutually conditionally independent** given Y if

$$P(X_1, X_2, \ldots, X_d \mid Y) = P(X_1 \mid Y) \cdot P(X_2 \mid Y) \cdots P(X_d \mid Y)$$



CSE 151A

Intro to Machine Learning

Lecture 05 – Part 02

How Conditional Independence Helps

Recall: The Bayes Classifier

- To use the Bayes classifier, we must estimate

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

for each class y_i, where $\vec{X} = (X_1, X_2, \ldots, X_d)$.

- Written differently, we need to estimate:

$$P(X_1 = x_1, \ldots, X_d = x_d \mid Y = y_i)$$

Recall: Histogram Estimators

- When $X_1, \dots, X_d$ are continuous, we can use **histogram estimators**.
- **Curse of Dimensionality**: if we discretize each dimension into 10 bins, there are 10^d bins.

Conditional Independence to the Rescue

- Now suppose $X_1, \ldots, X_d$ are mutually conditionally independent given Y. Then:

$$P(X_1 = x_1, \ldots, X_d = x_d \mid Y = y_i) = P(X_1 = x_1 \mid Y = y_i)P(X_2 = x_2 \mid Y = y_i) \cdots P(X_d = x_d \mid Y = y_i)$$

- Instead of estimating $P(X_1, \ldots, X_d \mid Y)$, estimate $P(X_1 \mid Y), \ldots, P(X_d \mid Y)$ separately.

Breaking the Curse

- Lets use histogram estimators.
- If we discretize each dimension into 10 bins, we need:
  - 10 bins to estimate $P(X_1|Y)$
  - 10 bins to estimate $P(X_2|Y)$
  - ...
  - 10 bins to estimate $P(X_d|Y)$
- We therefore need $10d$ bins in total.

Breaking the Curse

- Conditional independence **drastically reduced** the number of bins needed to cover the input space.
- From $\Theta(10^d)$ to $\Theta(d)$.

Idea

- Bayes Classifier needs a lot of data when d is big.
- But if the features are conditionally independent given the label, we don't need so much data.
- So let's just **assume** conditional independence.
- The result: the **Naïve Bayes Classifier**.

Naïve Bayes: The Algorithm

- **Assume** that $X_1, \ldots, X_d$ are mutually independent given the class label.
- Estimate $P(X_1 = x_1 \mid Y = y_i)$, ..., $P(X_d = x_d \mid Y = y_i)$ however you'd like: histograms, fitting univariate Gaussians, etc.
- Pick the y_i which maximizes

$$P(X_1 = x_1 \mid Y = y_i) \cdots P(X_d = x_d \mid Y = y_i) P(Y = y_i)$$

But wait...

- ...are we allowed to just assume conditional independence?

But wait...

- ...are we allowed to just assume conditional independence?
- **Answer:** who's going to stop us?

But wait...

- ...isn't the assumption wrong?

But wait...

- ...isn’t the assumption wrong?
- **Answer:** yeah, usually.

So does it even work?

So does it even work?

- **Answer:** Yes, surprisingly well.

“All models are wrong, but some are useful.”
- George Box, statistician

Estimating Probabilites

- You can estimate $P(X_i|Y)$ however makes sense.
- Often, people assume: **Gaussian Naïve Bayes**.

Example: NBA

- **Given**: player with height = 75 in, weight = 210 lbs.
- **Predict**: whether they are a forward or a guard.
- Let's use Gaussian Naïve Bayes.

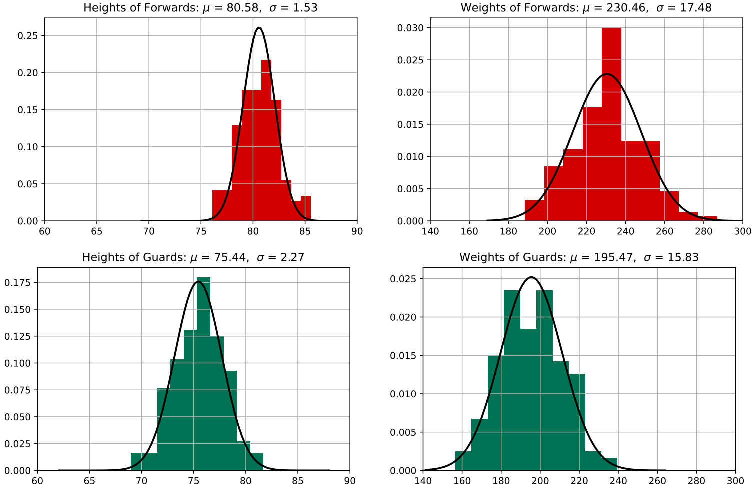
Example: NBA

- We need to estimate:

$$P(X_1 = 75 \mid Y = \text{forward})$$
$$P(X_1 = 75 \mid Y = \text{guard})$$
$$P(X_2 = 210 \mid Y = \text{forward})$$
$$P(X_2 = 210 \mid Y = \text{guard})$$

Example: NBA

- We'll fit 1-d Gaussians to:
    - heights of forwards.
    - heights of guards.
    - weights of forwards.
    - weights of guards.


Heights of Forwards: μ = 80.58, σ = 1.53
Weights of Forwards: μ = 230.46, σ = 17.48
Heights of Guards: μ = 75.44, σ = 2.27
Weights of Guards: μ = 195.47, σ = 15.83
0.00 0.05 0.10 0.15 0.20 0.25
60 65 70 75 80 85 90
0.000 0.005 0.010 0.015 0.020 0.025 0.030
140 160 180 200 220 240 260 280 300
0.000 0.025 0.050 0.075 0.100 0.125 0.150 0.175
60 65 70 75 80 85 90
0.000 0.005 0.010 0.015 0.020 0.025
140 160 180 200 220 240 260 280 300

Example: NBA

$$
\begin{aligned}
&P(X_1 = 75 \mid Y = \text{forward}) \cdot P(X_2 = 210 \mid Y = \text{forward}) \cdot P(Y = \text{forward}) \\
&\quad = \mathcal{N}(75; 80.58, 1.53^2) \cdot \mathcal{N}(210; 230.46, 17.48^2) \cdot \frac{156}{300} \\
&\quad \approx 6.73 \times 10^{-6}
\end{aligned}
$$

$$
\begin{aligned}
&P(X_1 = 75 \mid Y = \text{guard}) \cdot P(X_2 = 210 \mid Y = \text{guard}) \cdot P(Y = \text{guard}) \\
&\quad = \mathcal{N}(75; 75.44, 2.27^2) \cdot \mathcal{N}(210; 195.47, 15.83^2) \cdot \frac{144}{300} \\
&\quad \approx 5.88 \times 10^{-5}
\end{aligned}
$$

Example: NBA

- About 85% accurate on test set.
- But heights and weights are definitely not conditionally independent given position.

Gaussian Naïve Bayes

- $P(X_1 \mid Y) \cdots P(X_d \mid Y)$ is a product of 1-d Gaussians with different means, variances.

- Remember: result is a d-dimensional Gaussian with diagonal covariance matrix:

$$C = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \sigma_d^2 \end{pmatrix}$$

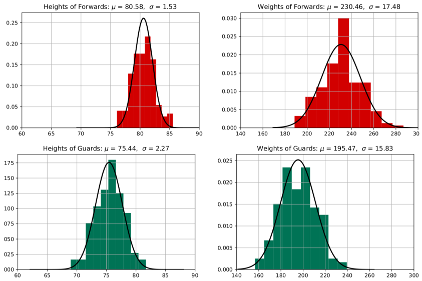
Gaussian Naïve Bayes

- But in GNB, each class has own diagonal covariance matrix.

- Therefore: Gaussian Naïve Bayes is **equivalent** to QDA with diagonal covariances.

Up next...

...predicting who survives on the Titanic.



CSE 151A

Intro to Machine Learning

Lecture 05 – Part 03

The Titanic

The Titanic Dataset

PassengerID	Survived	Pclass	Sex	Age	Fare	Embarked	FavColor
0	0	3	female	23.0	7.9250	S	yellow
1	0	1	male	47.0	52.0000	S	purple
2	0	3	male	36.0	7.4958	S	green
3	0	3	male	31.0	7.7500	Q	purple
4	0	3	male	19.0	7.8958	S	purple
...	...	...	...	...	...	...	...

Goal: predict survival given Age, Sex, Pclass.

Let's use Naïve Bayes

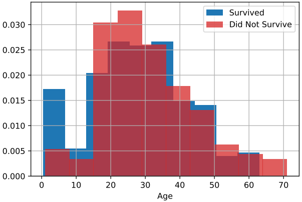
- We'll pick y_i so as to maximize

$$P(\text{Age} = x_1 \mid Y = y_i) \cdot P(\text{Sex} = x_2 \mid Y = y_i) \cdot P(\text{Pclass} = x_3 \mid Y = y_i) \cdot P(Y = y_i)$$

- We must choose how to estimate probabilities. Gaussians?

Estimating Probabilities

- How do we estimate $P(\text{Age} = x_1 \mid Y = y_i)$?
- Age is a continuous variable.
- Looks kind of bell-shaped, we'll fit Gaussians.



Estimating Probabilities

- How do we estimate $P(\text{Sex} = x_1 \mid Y = y_i)$?
- Sex is a **categorical** variable: either `male` or `female`.
- Fitting Gaussian makes no sense.
- But estimating these probabilities is easy.

Estimating Probabilities

$$P(\text{Sex = male} \mid \text{Survived}) \approx \frac{\text{\# of survived and male}}{\text{\# of survived}}$$

$$= .4$$

$$P(\text{Sex = male} \mid \text{Did Not Survive}) \approx \frac{\text{\# of died and male}}{\text{\# of died}}$$

$$= .87$$

Estimating Probabilities

- Pclass, too, is categorical. Estimate in same way.
- You can estimate $P(X_i|Y)$ however makes sense.
- **Can use different ways for different features**.
- Gaussian for age, simple ratio of counts for class, sex.

Example: The Titanic

- Using just age, sex, ticket class, Naïve Bayes is 70% accurate on test set.
- Not bad. Not great.
- To do better, add more features.

In High Dimensions

- Naïve Bayes excels in high dimensions.

- Example: document classification.
  - Document represented by a “bag of words”.
  - Pick a large number of words; say, 20,000.
  - Make a d-dimensional vector with ith entry counting number of occurrences of ith word.

Practical Issues

- We are multiplying lots of small probabilities:

$$P(X_1 | Y) \cdots P(X_d | Y)$$

- Potential for **underflow**.

Practical Issues

- "Trick": work with log-probabilities instead.
- Pick the y_i which maximizes

$$\log\left[P(X_1 = x_1 \mid Y = y_i) \cdots P(X_d = x_d \mid Y = y_i) P(Y = y_i)\right]$$

$$= \log P(X_1 = x_1 \mid Y = y_i) + \ldots + \log P(X_d = x_d \mid Y = y_i) + \log P(Y = y_i)$$

$$= \left(\sum_{j=1}^{d} \log P(X_j = x_j \mid Y = y_i)\right) + \log P(Y = y_i)$$