DSC 140A

Probabilistic Modeling & Machine Learning

Lecture 01 | Part 1

Recap

Applying the Bayes Classifier

- Predict the class y which maximizes:

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

- We must **estimate** the density, p_X.

- Two approaches:
  1. Non-parametric (e.g., histograms)
  2. Parametric (e.g., fit Gaussian with MLE)

Curse of Dimensionality

- In practice, we have many features.
- This means $p_X(\vec{X} = \vec{x} \mid Y = y)$ is **high dimensional**.
- Non-parametric estimators do not do well in high dimensions due to the **curse of dimensionality**:
  - Data required grows exponentially with number of features.

Responses

- Parametric density estimation can fare better.
- However, it too can suffer from the curse.
- **Today,** a different approach: assume **conditional independence**.

DSC 140A

Probabilistic Modeling & Machine Learning

Lecture 01 | Part 2

What is Conditional Independence?

Remember: Independence

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

$$\mathbb{P}(A, B) = \mathbb{P}(A) \cdot \mathbb{P}(B).$$

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

$$\mathbb{P}(A \mid B) = \mathbb{P}(A)$$

[1] or $\mathbb{P}(B) = 0$

Informally

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

Example

$P(A) = 1/4$

$P(A|B) = 1/4$

$P(A,B) = 4/52$

$P(A) = \frac{1}{4}$ $P(B) = \frac{16}{52}$

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?

Yes

♥: 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

$P(A,B) = 4/51$

$P(A) = \frac{13}{51}$ $P(B) = 15/51$

Example

$\frac{4}{51} \stackrel{?}{=} \frac{13}{51} \cdot \frac{15}{51}$

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? No

$P(B) = 15/51$

$P(B|A) = 4/13$

♥: 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...

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

In the Real World...

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

In the Real World...

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

In the Real World...

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

In the Real World...

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

Remember: Conditional Independence

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

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

- Equivalently[2]:

$$\mathbb{P}(A \mid B, C) = \mathbb{P}(A \mid C)$$

[2] Or $\mathbb{P}(B) = 0$

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*.

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

Example

$P(A|B,C) = P(A|C)$

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

- But they're (close) to **conditionally independent** given ticket price:
  - $\mathbb{P}(\text{Survived} = 1 \mid \text{PClass} = 1, \text{Fare} > 50) = .708$
  - $\mathbb{P}(\text{Survived} = 1 \mid \text{Fare} > 50) = .696$

More Variables

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

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

Densities

they are independent
if

- If A and B are **continuous** random variables, their joint density can be factored:

$$p(a, b) = p_A(a) \cdot p_B(b)$$

- If A and B are **conditionally independent** given C, then:

$$p(a, b \mid C = c) = p_A(a \mid C = c) \cdot p_B(b \mid C = c)$$

Densities

- Suppose $X_1, \ldots, X_d$ are d features, Y is class label.
- If the features are not independent given Y, then:

$$p(\vec{x} \mid Y = y) = p(x_1, x_2, \ldots, x_d \mid Y = y)$$

- **Curse of dimensionality**!

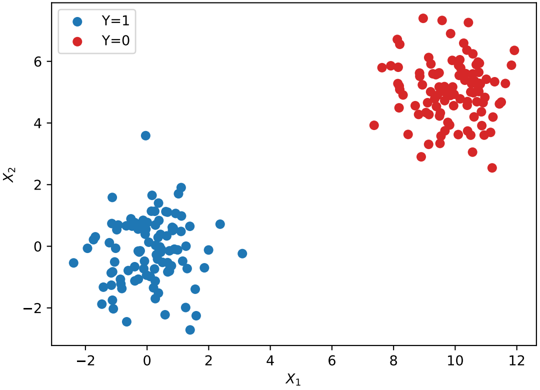
Densities

- Suppose $X_1, \ldots, X_d$ are d features, Y is class label.
- However, if the features are **mutually conditionally independent** given Y, then:

$$\begin{aligned} p(\vec{x} \mid Y = y) &= p(x_1, x_2, \ldots, x_d \mid Y = y) \\ &= p_1(x_1 \mid Y = y) \cdot p_2(x_2 \mid Y = y) \cdots p_d(x_d \mid Y = y) \end{aligned}$$

Exercise

Are X_1 and X_2 (close to) conditionally independent given Y? Yes



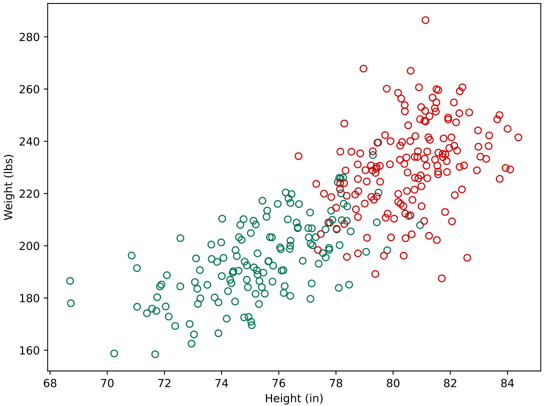
Exercise

Are height and weight (close to) conditionally independent given the player's position? No

$P(H, | W, \text{Green})$

$=$

$P(H | \text{Green})$



DSC 140A

Probabilistic Modeling & Machine Learning

Lecture 01 | Part 3

How Conditional Independence Helps

Recall: The Bayes Classifier

- To use the Bayes classifier, we must estimate

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

for each class y_i, where $\vec{x} = (x_1, x_2, \ldots, x_d)$.

- Written differently, we need to estimate:

$$p(x_1, \ldots, x_d \mid Y = y_i)$$

Recall: Histogram Estimators

- When $X_1, \ldots, X_d$ 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:

$$p(x_1, \ldots, x_d \mid Y = y_i) = p_1(x_1 \mid Y = y_i) p_2(x_2 \mid Y = y_i) \cdots p_d(x_d \mid Y = y_i)$$

- Instead of estimating $p(x_1, \ldots, x_d \mid Y)$, estimate $p_1(x_1 \mid Y), \ldots, p_d(x_d \mid Y)$ separately.

Breaking the Curse

- Suppose we use histogram estimators.

- If we discretize each dimension into 10 bins, we need:
  - 10 bins to estimate $p_1(x_1|Y)$
  - 10 bins to estimate $p_2(x_2|Y)$
  - ...
  - 10 bins to estimate $p_d(x_d|Y)$

- We therefore need $10d$ bins in total.

10^d vs $10d$

Breaking the Curse

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

$p(x \mid Y=y)\, P(Y=y)$

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**.

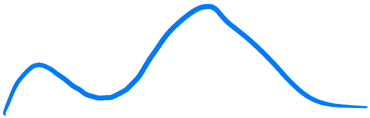
Naïve Bayes: The Algorithm

- **Assume** that $X_1, \dots, X_d$ are mutually independent given the class label.
- Estimate **one-dimensional** densities $p_1(x_1 \mid Y = y_i), \dots, p_d(x_d \mid Y = y_i)$ however you'd like.
  - histograms, fitting univariate Gaussians, etc.
- Pick the y_i which maximizes

$$p_1(x_1 \mid Y = y_i) \cdots p_2(x_d \mid Y = y_i)\mathbb{P}(Y = y_i)$$

But wait...

- ...are we allowed to just **assume** conditional independence?
- Sure!
- The independence assumption is usually **wrong**, but it can work surprisingly well in practice.



Estimating Probabilites

- You can estimate $p(X_i|Y)$ however makes sense.
- Popular: **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

- Compute:

$$p(75 \text{ in}, 210 \text{ lbs} \mid Y = \text{forward})\mathbb{P}(Y = \text{forward})$$

$$p(75 \text{ in}, 210 \text{ lbs} \mid Y = \text{guard})\mathbb{P}(Y = \text{guard})$$

- Using conditional independence assumption:

$$p_1(75 \text{ in} \mid Y = \text{forward}) \cdot p_2(210 \text{ lbs} \mid Y = \text{forward})\mathbb{P}(Y = \text{forward})$$

$$p_1(75 \text{ in} \mid Y = \text{guard}) \cdot p_2(210 \text{ lbs} \mid Y = \text{guard})\mathbb{P}(Y = \text{guard})$$

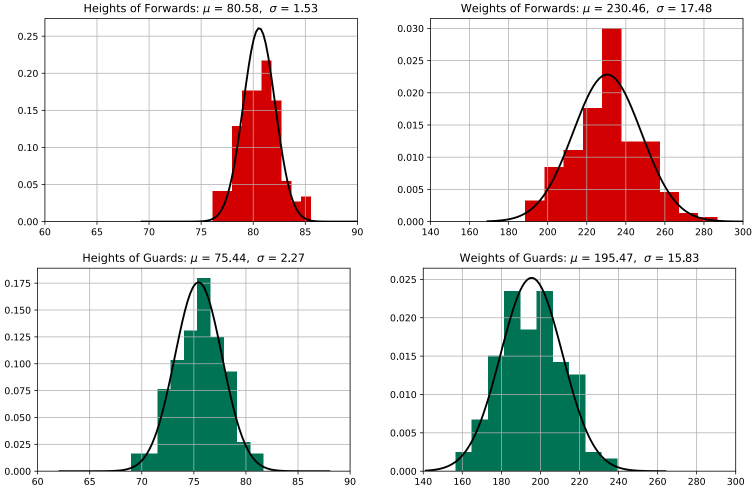
Example: NBA

- We need to estimate:

$$p_1(75 \text{ in} \mid Y = \text{forward})$$
$$p_1(75 \text{ in} \mid Y = \text{guard})$$
$$p_2(210 \text{ lbs} \mid Y = \text{forward})$$
$$p_2(210 \text{ lbs} \mid Y = \text{guard})$$

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

Example: NBA

$$
\begin{aligned}
&p_1(75 \mid Y = \text{forward}) \cdot p_2(210 \mid Y = \text{forward}) \cdot \mathbb{P}(Y = \text{forward}) \\
&\qquad = \mathcal{N}(75;\ 80.58,\ 1.53^2) \cdot \mathcal{N}(210;\ 230.46,\ 17.48^2) \cdot \frac{156}{300} \\
&\qquad \approx 6.73 \times 10^{-6}
\end{aligned}
$$

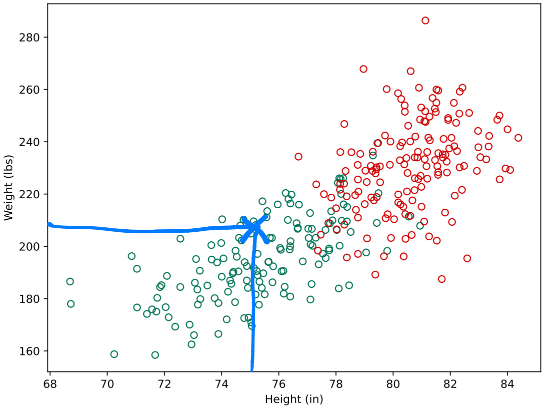
$$
\begin{aligned}
&p_1(75 \mid Y = \text{guard}) \cdot p_2(210 \mid Y = \text{guard}) \cdot \mathbb{P}(Y = \text{guard}) \\
&\qquad = \mathcal{N}(75;\ 75.44,\ 2.27^2) \cdot \mathcal{N}(210;\ 195.47,\ 15.83^2) \cdot \frac{144}{300} \\
&\qquad \approx 5.88 \times 10^{-5}
\end{aligned}
$$

Example: NBA

- About 85% accurate on test set.

Exercise

Are height and weight conditionally independent given the player's position? No



Example: NBA

- No!

- Gaussian Naïve Bayes worked well even though the conditional independence assumption is not accurate.

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.

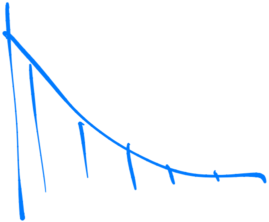
- Remember: result is a d-dimensional Gaussian with diagonal covariance matrix:

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

Gaussian Naïve Bayes

- But in GNB, each class has own diagonal covariance matrix.
- Therefore: Gaussian Naïve Bayes is **equivalent** to QDA with diagonal covariances.

Beyond Gaussian



- Naïve Bayes is very flexible.
- Can use different parametric distributions for different features.
  - E.g., normal for feature 1, log normal for feature 2, etc.
- Can use non-parametric density estimation (densities) for other features.
- Can also handle discrete features.

Up next…

…predicting who survives on the Titanic.

DSC 140A

Probabilistic Modeling & Machine Learning

Lecture 01 | Part 4

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

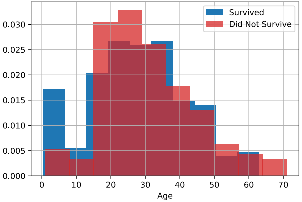
- We'll pick y_i so as to maximize

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

- 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 $\mathbb{P}(\text{Sex} = x_1 \mid Y = y_i)$?
- Sex is a **discrete** variable in this data set.
- Fitting Gaussian makes no sense.
- But estimating these probabilities is easy.

Estimating Probabilities

$$\mathbb{P}(\text{Sex} = \text{male} \mid \text{Survived}) \approx \frac{\#\text{ of survived and male}}{\#\text{ of survived}} = .4$$

$$\mathbb{P}(\text{Sex} = \text{male} \mid \text{Did Not Survive}) \approx \frac{\#\text{ of died and male}}{\#\text{ of died}} = .87$$

Estimating Probabilities

- Pclass, too, is categorical. Estimate in same way.
- You can estimate $\mathbb{P}(X_i | Y)$ 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 can work well 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:

$$\mathbb{P}(X_1 \mid Y) \cdots \mathbb{P}(X_d \mid Y)$$

- Potential for **underflow**.

Practical Issues

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

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

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

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