Naïve Bayes classifier

Introduction

- Review of Conditional Probability
- Naïve Bayes Classifier

Example Dataset

Consider a dataset of weather conditions and whether to play tennis

Outlook	Temperature	Humidity	Wind	PlayTennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rainy	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes

Conditional Probability

- Definition: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Conditional Probability

- Definition: $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- Example: Probability of playing tennis given that it's sunny

$$P(\text{PlayTennis=Yes}|\text{Outlook=Sunny})$$

Calculating Conditional Probability

$$P(\text{PlayTennis=Yes}|\text{Outlook=Sunny})$$

- Total instances: 13

Calculating Conditional Probability

$$P(\text{PlayTennis=Yes}|\text{Outlook=Sunny})$$

- Total instances: 13
- Instances where Outlook=Sunny: 5

Calculating Conditional Probability

$$P(\text{PlayTennis=Yes}|\text{Outlook=Sunny})$$

- Total instances: 13
- Instances where Outlook=Sunny: 5
- Instances where PlayTennis=Yes and Outlook=Sunny: 2

Calculating Conditional Probability

$$P(\text{PlayTennis=Yes}|\text{Outlook=Sunny})$$

- Total instances: 13
- Instances where Outlook=Sunny: 5
- Instances where PlayTennis=Yes and Outlook=Sunny: 2
- $P(\text{PlayTennis=Yes}|\text{Outlook=Sunny}) =$

Naïve Bayes Classifier

Steps:

1. Estimate conditional probabilities:

$$P(Y = \text{Yes}|X_1, X_2, \ldots)$$
$$P(Y = \text{No}|X_1, X_2, \ldots)$$

Naïve Bayes Classifier

Steps:

1. Estimate conditional probabilities:

$$P(Y = \text{Yes}|X_1, X_2, \ldots)$$
$$P(Y = \text{No}|X_1, X_2, \ldots)$$

2. Make prediction:

$$\widehat{Y} = \text{Yes if } P(Y = \text{Yes}|X_1, X_2, \ldots) \geq P(Y = \text{No}|X_1, X_2, \ldots)$$
$$\widehat{Y} = \text{No if } P(Y = \text{Yes}|X_1, X_2, \ldots) < P(Y = \text{No}|X_1, X_2, \ldots)$$

Naïve Bayes Classifier

- Example:
  - Outlook $=$ Sunny, Temperature $=$ Cool, Humidity $=$ High, Wind $=$ Strong
  - What is the probability that PlayTennis $=$ Yes?

Naïve Bayes Classifier

- Example:
  - Outlook $=$ Sunny, Temperature $=$ Cool, Humidity $=$ High, Wind $=$ Strong
  - What is the probability that PlayTennis $=$ Yes?

- Counting from the table yields

$$P(\text{PlayTennis} = \text{Yes}|\text{Sunny, Cool, High, Strong}) = 0$$

But this is probably not accurate!

Naïve Bayes Classifier

- Instead, we use Bayes' Theorem:

$$P(Y|X_1, X_2, \ldots) = \frac{P(X_1, X_2, \ldots |Y)P(Y)}{P(X_1, X_2, \ldots)}$$

Naïve Bayes Classifier

- Instead, we use Bayes' Theorem:

$$P(Y|X_1, X_2, \ldots) = \frac{P(X_1, X_2, \ldots|Y)P(Y)}{P(X_1, X_2, \ldots)}$$

- $\ldots$ and assumes *conditional* independence between predictors

$$P(X_1, X_2, \ldots|Y) = P(X_1|Y)P(X_2|Y)\ldots$$

- Now $P(X_1|Y), P(X_2|Y), \ldots$ can be accurately estimated with only a few instances!

Prediction

We will make prediction $\widehat{Y} =$ Yes if

$$P(Y = \text{Yes}|X_1, X_2, \ldots) > P(Y = \text{No}|X_1, X_2, \ldots),$$

and vice versa

Prediction

We will make prediction $\widehat{Y} = \text{Yes}$ if

$$P(Y = \text{Yes}|X_1, X_2, \ldots) > P(Y = \text{No}|X_1, X_2, \ldots),$$

and vice versa

$$P(Y = \text{Yes}|X_1, X_2, \ldots) = \frac{P(X_1, X_2, \ldots|Y = \text{Yes})P(Y = \text{Yes})}{P(X_1, X_2, \ldots)}$$

$$P(Y = \text{No}|X_1, X_2, \ldots) = \frac{P(X_1, X_2, \ldots|Y = \text{No})P(Y = \text{No})}{P(X_1, X_2, \ldots)}$$

Prediction

To compare these two, we do not need to compute $P(X_1, X_2, \ldots)$

We predict $\widehat{Y} =$ Yes if

$$\begin{aligned} &P(X_1, X_2, \ldots | Y = \text{Yes}) \times P(Y = \text{Yes}) \\ > &P(X_1, X_2, \ldots | Y = \text{No}) \times P(Y = \text{No}) \end{aligned}$$

Prediction

To compare these two, we do not need to compute $P(X_1, X_2, \ldots)$

With conditional independence, we predict $\widehat{Y} =$ Yes if

$$
\begin{aligned}
&P(X_1|Y = \text{Yes}) \times P(X_2|Y = \text{Yes}) \times \ldots \times P(Y = \text{Yes}) \\
>&P(X_1|Y = \text{No}) \times P(X_2|Y = \text{No}) \times \ldots \times P(Y = \text{No})
\end{aligned}
$$

and vice versa

Example Dataset

Dataset of weather conditions and whether to play tennis

Outlook	Temperature	Humidity	Wind	PlayTennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rainy	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes

Predicting with Naïve Bayes

- Example: Predict PlayTennis given {Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=Strong}

Predicting with Naïve Bayes

- Example: Predict PlayTennis given {Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=Strong}
- Calculate posterior probabilities for both classes (Yes and No)
- $P(\text{Yes}|\text{data}) = P(\text{Outlook=Sunny}|\text{Yes}) \times P(\text{Temperature=Cool}|\text{Yes}) \times \ldots \times P(\text{Yes})$
- $P(\text{No}|\text{data}) = P(\text{Outlook=Sunny}|\text{No}) \times P(\text{Temperature=Cool}|\text{No}) \times \ldots \times P(\text{No})$

Calculating Priors and Likelihoods

- Priors:
  - $P(\text{Yes}) = \frac{9}{13}$
  - $P(\text{No}) = \frac{5}{13}$
- Likelihoods:
  - $P(\text{Outlook=Sunny}|\text{Yes}) = \frac{2}{9}$
  - $P(\text{Outlook=Sunny}|\text{No}) = \frac{3}{4}$
  - And similarly for other features

Final Prediction

- Compare $P(\text{Yes}|\text{data})$ and $P(\text{No}|\text{data})$
- Predict the class with the higher posterior probability
- In this example:
  $P(\text{Yes}|\text{data}) < P(\text{No}|\text{data}) \Rightarrow \text{PlayTennis} = \text{No}$

Continuous Features

- We can also handle continuous features with Naïve Bayes
- Assume the continuous values follow a Gaussian (normal) distribution
- Use Gaussian likelihood for these features

Example Dataset

A dataset of student performance.

Study Hours	Previous Grade	Pass
1.5	C	No
3.0	B	Yes
2.0	C	No
4.0	A	Yes
2.5	B	No
3.5	A	Yes
3.0	C	Yes
5.0	A	Yes
1.0	C	No
4.5	B	Yes

Gaussian Naïve Bayes

- For a continuous feature x, likelihood is given by:

$$P(x|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x-\mu_y)^2}{2\sigma_y^2}\right)$$

where μ_y and σ_y^2 are the mean and variance of the feature for class y

Example: Continuous Feature

- Consider Study Hours as a continuous feature
- Calculate mean (μ) and variance (σ^2) for each class (Pass = Yes, No)

$$\mu_{\text{Yes}} = \frac{\sum \text{Study Hours for Yes}}{N_{\text{Yes}}}$$

$$\sigma^2_{\text{Yes}} = \frac{\sum (\text{Study Hours for Yes} - \mu_{\text{Yes}})^2}{N_{\text{Yes}} - 1}$$

Calculating Parameters

- For Pass = Yes:
  - $\mu_{\text{Yes}} = \frac{3.0+4.0+3.5+5.0+3.0+4.5}{6} = 3.67$
  - $\sigma^2_{\text{Yes}} = \frac{(3.0-3.67)^2+(4.0-3.67)^2+...}{5} = 0.73$
- For Pass = No:
  - $\mu_{\text{No}} = \frac{1.5+2.0+2.5+1.0}{4} = 1.75$
  - $\sigma^2_{\text{No}} = \frac{(1.5-1.75)^2+(2.0-1.75)^2+...}{3} = 0.58$

Prediction with Continuous Feature

- Example: Predict Pass given {Study Hours=3.2, Previous Grade=B}
- Use Gaussian likelihood for Study Hours
- Calculate $P(\text{Study Hours} = 3.2|\text{Yes})$ and $P(\text{Study Hours} = 3.2|\text{No})$

Posterior Probability with Continuous Feature

- Compute the posterior probabilities:

$$P(\text{Yes}|\text{data}) \\ \approx P(\text{Study Hours = 3.2}|\text{Yes}) \times P(\text{Previous Grade = B}|\text{Yes}) \times P(\text{Yes})$$

$$P(\text{No}|\text{data}) \\ \approx P(\text{Study Hours = 3.2}|\text{No}) \times P(\text{Previous Grade = B}|\text{No}) \times P(\text{No})$$

- Compare and predict the class with higher posterior probability

Computing Likelihoods

$\mu_{\text{Yes}} = 3.67, \sigma^2_{\text{Yes}} = 0.73, \mu_{\text{No}} = 1.75, \sigma^2_{\text{No}} = 0.58$

$P(\text{Study Hours = 3.2}|\text{Yes}) =$

$P(\text{Study Hours = 3.2}|\text{No}) =$

Computing Likelihoods

$$P(\text{Previous Grade = B}|\text{Yes}) =$$

$$P(\text{Previous Grade = B}|\text{No}) =$$

Positive or negative movie review?

- This movie is disappointing.
- I love everything about this movie.
- I would love to have that two hours of my life back.
- This is one of my favorite if not favorite films.
- I have seen so many bad low budget movies lately, but I love this one.

Naïve Bayes for text

$$P(w_1, w_2, \ldots, w_n|y)P(y) = P(w_1|y)P(w_2|y)\ldots P(w_n|y)P(y)$$

where

$$P(w_i|y) = \frac{count(w_i, y)}{\sum_{w \in V} count(w, y)}$$

and

$$P(y) = \frac{countdoc(Y = y)}{count(\text{Documents})}$$

Example

- This movie is disappointing.
- I love everything about this movie.
- I would love to have that two hours of my life back.
- This is one of my favorite if not favorite films.
- I have seen so many bad low budget movies lately, but I love this one.

$$P(\text{Positive}) =$$

Example

- This movie is disappointing.
- I love everything about this movie.
- I would love to have that two hours of my life back.
- This is one of my favorite if not favorite films.
- I have seen so many bad low budget movies lately, but I love this one.

$$P(\text{favorite}|\text{Positive}) =$$

Example

$$P(y = 1 | \text{I love, love this movie.})$$

Example

$$P(y = 0 | \text{I love, love this movie.})$$

Laplace smoothing

- Want to predict the class of "I **slept** through the entire movie" but the word **slept** is not in the training set

$$P(\textsf{slept}|y) = \frac{count(\textsf{slept}, y)}{\sum_{w \in V} count(w, y)} = 0.$$

Laplace smoothing

- Want to predict the class of "I **slept** through the entire movie" but the word **slept** is not in the training set

$$P(\textsf{slept}|y) = \frac{count(\textsf{slept}, y)}{\sum_{w \in V} count(w, y)} = 0.$$

- There is no best y in this case.

$$P(y|\textsf{slept}, \ldots) = P(\textsf{slept}|y) \times \ldots \times P(y) = 0$$

Laplace smoothing

Fix $\alpha > 0$.

$$P(w_i|y) = \frac{count(w_i, y) + \alpha}{\sum_{w \in V}(count(w, y) + \alpha)}$$

$$= \frac{count(w_i, y) + \alpha}{\sum_{w \in V} count(w, y) + \alpha|Vocab|}$$

Laplace smoothing

Fix $\alpha > 0$.

$$P(w_i|y) = \frac{count(w_i, y) + \alpha}{\sum_{w \in V}(count(w, y) + \alpha)}$$

$$= \frac{count(w_i, y) + \alpha}{\sum_{w \in V} count(w, y) + \alpha|Vocab|}$$

For example, if we choose $\alpha = 1$,

$$P(\textsf{slept}|y) = \frac{1}{\sum_{w \in V} count(w, y) + |Vocab|} \neq 0.$$

Learning Naïve Bayes

- From the training corpus, extract the **Vocabulary**.
- For each class y, calculate $P(y)$
  - Count number of documents in class y.
  - $$P(y) = \frac{countdoc(Y=y)}{count(\text{Documents})}$$
- For each word w_i and class y
  - Merge all documents in class y
  - $n_i \leftarrow$ # of occurrence of each word in class y
  - $$P(w_i|y) = \frac{n_i+\alpha}{\sum_i n_i+\alpha|Vocab|}$$