

CSE422: Artificial intelligence

Naïve-Bayes Classifier

**Asif Shahriar
Lecturer, CSE, BRACU**

BAYES' Rule

- From the formula of conditional probability $P(x | y) = P(x, y) / P(y)$, we can infer:
 $P(x, y) = P(x | y) \times P(y)$ [called the product rule]
- Similarly, $P(x, y) = P(y | x) \times P(x)$
- Dividing we get:

$$P(x|y) = \frac{P(y|x)}{P(y)} P(x)$$

- In other words: $P(\text{cause}|\text{effect}) = \frac{P(\text{effect}|\text{cause})P(\text{cause})}{P(\text{effect})}$
- Why is this SO important?
 - It is hard to know the cause, but it is easy to see the effect
 - Using this formula, we can infer the cause by analyzing the effect

BAYES' Rule

$$P(x|y) = \frac{\text{Likelihood}}{\text{Predictor-prior probability}} P(y|x) P(x)$$

Posterior probability

Likelihood

Predictor-prior probability

Class-prior probability

BAYES' Rule

| Temperature (F) | Play Tennis |
|-----------------|-------------|
| 70 | Yes |
| 32 | No |
| 65 | No |
| 75 | Yes |
| 30 | No |
| 75 | Yes |
| 72 | No |

- Consider two boolean RVs, A: “We will play tennis” and B: “Warm day ($\text{Temp} \geq 50 \text{ F}$)”
 - Think ML: Temperature is “**feature**”, Play tennis is “**target**”
- If it is a warm day, what is the probability that we will play tennis?
 - Think ML: Given a **feature**, predict the **target**

BAYES' Rule

| Temperature (F) | Play Tennis |
|-----------------|-------------|
| 70 | Yes |
| 32 | No |
| 65 | No |
| 75 | Yes |
| 30 | No |
| 75 | Yes |
| 72 | No |

- If it is a warm day, what is the probability that we will play tennis?
- $P(b | a) = 1$
- $P(a) = 3/7$
- $P(b) = 5/7$
- Therefore, $P(a | b) = 0.6$
 - Think ML: **Output = YES**

Why BAYES is DIFFICULT in Practice

| Wind | Humidity | Temperature | Play Tennis |
|------|----------|-------------|-------------|
| 5 | 95 | 70 | Yes |
| 10 | 80 | 32 | No |
| 20 | 80 | 65 | No |
| 10 | 85 | 75 | Yes |
| 8 | 35 | 30 | No |
| 8 | 35 | 75 | Yes |
| 25 | 35 | 72 | No |

- Now let there are FOUR boolean RVs: A (“Play tennis”), B(“Warm day”), C(“Dry day”), D(“Windy day”) [Think ML: Features: B, C, D, Target: A]
- We want to find the probability of playing tennis in a warm, dry, windy day $P(a | b, c, d)$
- For this we need the likelihood $P(b, c, d | a)$
- Computing this is REALLY difficult and makes it hard to use BAYES in real applications
 - For this problem it is actually easy (do it)
 - But this is a small table with 7 entries only
 - In real world, to compute such a likelihood with good accuracy would require an extensive record with millions of entries – very expensive and difficult

Naïve-Bayes Classifier

- Computing the likelihood $P(b, c, d | a)$ is difficult in real world applications
- However, we can APPROXIMATE the calculation, making it practical for real use
- The **Naïve assumption**: All **features** are **conditionally independent** of each other given the target label
- In other words, B, C, D are conditionally independent given A
- So, using our formula for conditional independence, we get:
- $P(b, c, d | a) = P(b | a) \times P(c | a) \times P(d | a)$
- This is much easier to calculate from records!
- The Naïve Bayes classifier often does surprisingly well, outperforming more sophisticated classification methods
- $P(X_1, X_2, \dots, X_n | C) = P(X_1 | C)P(X_2 | C) \dots P(X_n | C)$

Approximation using Naïve-BAYES

| Wind | Humidity | Temperature | Play Tennis |
|------|----------|-------------|-------------|
| 5 | 95 | 70 | Yes |
| 10 | 80 | 32 | No |
| 20 | 80 | 65 | No |
| 10 | 85 | 75 | Yes |
| 8 | 35 | 30 | No |
| 8 | 35 | 75 | Yes |
| 25 | 35 | 72 | No |

- We want to find the probability of playing tennis in a warm, dry, windy day $P(a | b, c, d)$
- $P(a | b, c, d) = P(b, c, d | a) \times P(a) / P(b, c, d)$
- Do it yourself!

Naïve-BAYES Example

PlayTennis: training examples

| Day | Outlook | Temperature | Humidity | Wind | PlayTennis |
|-----|----------|-------------|----------|--------|------------|
| D1 | Sunny | Hot | High | Weak | No |
| D2 | Sunny | Hot | High | Strong | No |
| D3 | Overcast | Hot | High | Weak | Yes |
| D4 | Rain | Mild | High | Weak | Yes |
| D5 | Rain | Cool | Normal | Weak | Yes |
| D6 | Rain | Cool | Normal | Strong | No |
| D7 | Overcast | Cool | Normal | Strong | Yes |
| D8 | Sunny | Mild | High | Weak | No |
| D9 | Sunny | Cool | Normal | Weak | Yes |
| D10 | Rain | Mild | Normal | Weak | Yes |
| D11 | Sunny | Mild | Normal | Strong | Yes |
| D12 | Overcast | Mild | High | Strong | Yes |
| D13 | Overcast | Hot | Normal | Weak | Yes |
| D14 | Rain | Mild | High | Strong | No |

Step 1 – Learning: From the training data, we will **LEARN** the **conditional probability** distribution of each **Feature** (Outlook, Temperature, Humidity, Wind) Given the **Target** (PlayTennis)

Naïve-BAYES Example

PlayTennis: training examples

| Day | Outlook | Temperature | Humidity | Wind | PlayTennis |
|-----|----------|-------------|----------|--------|------------|
| D1 | Sunny | Hot | High | Weak | No |
| D2 | Sunny | Hot | High | Strong | No |
| D3 | Overcast | Hot | High | Weak | Yes |
| D4 | Rain | Mild | High | Weak | Yes |
| D5 | Rain | Cool | Normal | Weak | Yes |
| D6 | Rain | Cool | Normal | Strong | No |
| D7 | Overcast | Cool | Normal | Strong | Yes |
| D8 | Sunny | Mild | High | Weak | No |
| D9 | Sunny | Cool | Normal | Weak | Yes |
| D10 | Rain | Mild | Normal | Weak | Yes |
| D11 | Sunny | Mild | Normal | Strong | Yes |
| D12 | Overcast | Mild | High | Strong | Yes |
| D13 | Overcast | Hot | Normal | Weak | Yes |
| D14 | Rain | Mild | High | Strong | No |

Learning Phase

| Outlook | Play=Yes | Play=No |
|----------|----------|---------|
| Sunny | 2/9 | 3/5 |
| Overcast | 4/9 | 0/5 |
| Rain | 3/9 | 2/5 |

| Temperature | Play=Yes | Play=No |
|-------------|----------|---------|
| Hot | 2/9 | 2/5 |
| Mild | 4/9 | 2/5 |
| Cool | 3/9 | 1/5 |

| Humidity | Play=Yes | Play=No |
|----------|----------|---------|
| High | 3/9 | 4/5 |
| Normal | 6/9 | 1/5 |

| Wind | Play=Yes | Play=No |
|--------|----------|---------|
| Strong | 3/9 | 3/5 |
| Weak | 6/9 | 2/5 |

$$P(\text{Play}=\text{Yes}) = 9/14$$

$$P(\text{Play}=\text{No}) = 5/14$$

Caution: This is a **CONDITIONAL** probability table, NOT Joint Probability Distribution Table

Naïve-BAYES Example

Step 2 – **Testing:** Using what we learned (likelihoods) during training, we will **PREDICT** the label (target) for a new, unseen input data

Learning Phase

| Outlook | Play=Yes | Play=No |
|----------|----------|---------|
| Sunny | 2/9 | 3/5 |
| Overcast | 4/9 | 0/5 |
| Rain | 3/9 | 2/5 |

| Temperature | Play=Yes | Play=No |
|-------------|----------|---------|
| Hot | 2/9 | 2/5 |
| Mild | 4/9 | 2/5 |
| Cool | 3/9 | 1/5 |

| Humidity | Play=Yes | Play=No |
|----------|----------|---------|
| High | 3/9 | 4/5 |
| Normal | 6/9 | 1/5 |

$$P(\text{Play}=\text{Yes}) = 9/14$$

| Wind | Play=Yes | Play=No |
|--------|----------|---------|
| Strong | 3/9 | 3/5 |
| Weak | 6/9 | 2/5 |

$$P(\text{Play}=\text{No}) = 5/14$$

Test Phase

- Given a new instance, predict its label

$$\mathbf{x}' = (\text{Outlook}=\text{Sunny}, \text{Temperature}=\text{Cool}, \text{Humidity}=\text{High}, \text{Wind}=\text{Strong})$$

- Look up tables achieved in the learning phrase

$$P(\text{Outlook}=\text{Sunny} | \text{Play}=\text{Yes}) = 2/9$$

$$P(\text{Outlook}=\text{Sunny} | \text{Play}=\text{No}) = 3/5$$

$$P(\text{Temperature}=\text{Cool} | \text{Play}=\text{Yes}) = 3/9$$

$$P(\text{Temperature}=\text{Cool} | \text{Play}=\text{No}) = 1/5$$

$$P(\text{Humidity}=\text{High} | \text{Play}=\text{Yes}) = 3/9$$

$$P(\text{Humidity}=\text{High} | \text{Play}=\text{No}) = 4/5$$

$$P(\text{Wind}=\text{Strong} | \text{Play}=\text{Yes}) = 3/9$$

$$P(\text{Wind}=\text{Strong} | \text{Play}=\text{No}) = 3/5$$

$$P(\text{Play}=\text{Yes}) = 9/14$$

$$P(\text{Play}=\text{No}) = 5/14$$

- Decision making

$$P(\text{Yes} | \mathbf{x}') \approx [P(\text{Sunny} | \text{Yes})P(\text{Cool} | \text{Yes})P(\text{High} | \text{Yes})P(\text{Strong} | \text{Yes})]P(\text{Play}=\text{Yes}) = 0.0053$$

$$P(\text{No} | \mathbf{x}') \approx [P(\text{Sunny} | \text{No})P(\text{Cool} | \text{No})P(\text{High} | \text{No})P(\text{Strong} | \text{No})]P(\text{Play}=\text{No}) = 0.0206$$

Given the fact $P(\text{Yes} | \mathbf{x}') < P(\text{No} | \mathbf{x}')$, we label \mathbf{x}' to be “No”.

Naïve-BAYES for Continuous-valued Features

- Features are not always discrete, sometimes they are continuous
- Example: Temperature is a naturally continuous RV
 - Yes: 25.2, 19.3, 18.5, 21.7, 20.1, 24.3, 22.8, 23.1, 19.8
 - No: 27.3, 30.1, 17.4, 29.5, 15.1
- Problem: Very unlikely that we will find an exact value of temperature
 - E.g. $P(\text{Play} = \text{YES} | \text{Temp} = 25.8)$, but $\text{Temp} = 25.8$ DOES NOT EXIST in data
- Solution: Model continuous probabilities with **Normal Distribution**

$$\hat{P}(X_j | C = c_i) = \frac{1}{\sqrt{2\pi}\sigma_{ji}} \exp\left(-\frac{(X_j - \mu_{ji})^2}{2\sigma_{ji}^2}\right)$$

μ_{ji} : mean (average) of feature values X_j of examples for which $C = c_i$

σ_{ji} : standard deviation of feature values X_j of examples for which $C = c_i$

Naïve-BAYES for Continuous-valued Features

$$\hat{P}(X_j | C = c_i) = \frac{1}{\sqrt{2\pi}\sigma_{ji}} \exp\left(-\frac{(X_j - \mu_{ji})^2}{2\sigma_{ji}^2}\right)$$

μ_{ji} : mean (avearage) of feature values X_j of examples for which $C = c_i$

σ_{ji} : standard deviation of feature values X_j of examples for which $C = c_i$

Learning Phase: for $\mathbf{X} = (X_1, \dots, X_n)$, $C = c_1, \dots, c_L$

Output: $n \times L$ normal distributions and $P(C = c_i)$ $i = 1, \dots, L$

Test Phase: Given an unknown instance $\mathbf{X}' = (a'_1, \dots, a'_n)$

- Instead of looking-up tables, calculate conditional probabilities with all the normal distributions achieved in the learning phrase

Naïve-BAYES for Continuous-valued Features

- Let's consider the temperature example again
 - Yes: 25.2, 19.3, 18.5, 21.7, 20.1, 24.3, 22.8, 23.1, 19.8
 - No: 27.3, 30.1, 17.4, 29.5, 15.1
- Estimate mean and variance for each class

$$\mu = \frac{1}{N} \sum_{n=1}^N x_n, \quad \sigma^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu)^2$$

$$\mu_{Yes} = 21.64, \quad \sigma_{Yes} = 2.35$$

$$\mu_{No} = 23.88, \quad \sigma_{No} = 7.09$$

- Learning Phase:** output two Gaussian models for $P(\text{temp}|C)$

$$\hat{P}(x | Yes) = \frac{1}{2.35\sqrt{2\pi}} \exp\left(-\frac{(x - 21.64)^2}{2 \times 2.35^2}\right) = \frac{1}{2.35\sqrt{2\pi}} \exp\left(-\frac{(x - 21.64)^2}{11.09}\right)$$

$$\hat{P}(x | No) = \frac{1}{7.09\sqrt{2\pi}} \exp\left(-\frac{(x - 23.88)^2}{2 \times 7.09^2}\right) = \frac{1}{7.09\sqrt{2\pi}} \exp\left(-\frac{(x - 23.88)^2}{50.25}\right)$$

Relevant Issues to consider for Naïve-BAYES

- **Violation of Independence Assumption**

- For most real worlds tasks, $P(X_1, \dots, X_n | C) \neq P(X_1 | C) \cdots P(X_n | C)$ [i.e. features are not independent]
- Nonetheless, Naïve-BAYES works surprisingly well anyway
- Check this out for a comparison against other models: [notebook](#)

- **Zero conditional probability Problem**

- If no example contains the feature value $X_j = a_{jk}$, $\hat{P}(X_j = a_{jk} | C = c_i) = 0$
- In this circumstance, $\hat{P}(x_1 | c_i) \cdots \hat{P}(a_{jk} | c_i) \cdots \hat{P}(x_n | c_i) = 0$ during test
- For a remedy, conditional probabilities re-estimated with

$$\hat{P}(X_j = a_{jk} | C = c_i) = \frac{n_c + mp}{n + m}$$

n_c : number of training examples for which $X_j = a_{jk}$ and $C = c_i$

n : number of training examples for which $C = c_i$

p : prior estimate (usually, $p = 1/t$ for t possible values of X_j)

m : weight to prior (number of "virtual" examples, $m \geq 1$)