Machine Learning (Lab support)

# Naïve Bayes

**Abdelkrime Aries**
*Laboratoire de la Communication dans les Systèmes Informatiques (LCSI)*
*École nationale Supérieure d'Informatique (ESI, ex. INI), Algiers, Algeria*

**Academic year: 2024-2025**

# Machine Learning (Lab support)

## Naïve Bayes: Bayes theory

$$\overbrace{P(A|B)}^{\text{Posterior}} = \frac{\overbrace{P(A)}^{\text{Prior}}\ \overbrace{P(B|A)}^{\text{Likelihood}}}{\underbrace{P(B)}_{\text{Evidence}}}$$

# Machine Learning (Lab support)

## Naïve Bayes: Plan

1. **Theoretical formulation**
   - Estimation
   - Prior probability
   - Likelihood

2. **Numerical application**
   - Multinomial NB
   - Bernoulli NB
   - Normal NB

Section 1

# Theoretical formulation

# Naïve Bayes

## Theoretical formulation

- Given a sample $x$, the probability of generating a class $k$ can be expressed as:

$$p(y = k|x) = \frac{p(y = k)p(x|y = k)}{p(x)}$$

- Given $L$ classes, the output class is the one that maximizes this probability

$$\hat{y} = \arg\max_{k} p(y = k|x), \ k \in \{1, \cdots, L\}$$

- in this case, no need for Evidence probability (not dependent to $y$)

$$p(y = k|x) \propto p(y = k)p(x|y = k)$$

# Naïve Bayes: Theoretical formulation

## Estimation

- The output class $\hat{y}$ is estimated as
$$\hat{y} = \arg\max_{k} p(y=k|x) = \arg\max_{k} p(y=k)p(x|y=k),\ k \in \{1, \cdots, L\}$$
- **The naive part**: the assumption of features independence
$$p(x|y=k) \approx \prod_{j=1}^{N} P(x_j|y=k)$$
- In this case, the estimation function would be:
$$\hat{y} = \arg\max_{y=k} p(y=k) \prod_{j=1}^{N} P(x_j|y=k),\ k \in \{1, \cdots, L\}$$
- In practice, the calculation is simplified
$$\hat{y} = \arg\max_{y=k} \log p(y=k) + \sum_{j=1}^{N} \log p(x_j|y=k),\ k \in \{1, \cdots, L\}$$

# Naïve Bayes: Theoretical formulation

## Prior probability

$$p(y = k) = \frac{|\{y^{(i)} = k,\ i \in \{1, \cdots, M\}\}|}{M}$$

- $|\{y^{(i)} = k,\ i \in \{1, \cdots, M\}\}|$ is the number of training samples having $k$ as class
- $M$ is the size of the training dataset
- If classes' distribution is uniform, this probability can be ignored
- If we want to give the same prior probability to classes, this probability can be ignored

# Naïve Bayes: Theoretical formulation
## Likelihood: Multinomial distribution

$$p(x_j = v|y = k) = \frac{|\{y^{(i)} = k \wedge x_j^{(i)} = v,\ i \in \{1, \cdots, M\}\}|}{|\{y^{(i)} = k,\ i \in \{1, \cdots, M\}\}|} = \frac{\#(y = k \wedge x_j = v)}{\#(y = k)}$$

- $x_j$ is a categorical feature having a value $v$
- $v$ is a value among unique values $V_j$ (called ***vocabulary***) of the feature $j$
- $\#(y = k \wedge x_j = v)$ is the number of training samples with feature $j$ equals to $v$ and having $k$ as class
- $\#(y = k)$ is the number of training samples having $k$ as class
- Smoothing can be used in case there are unseen values $v$ in the test dataset, where $V_j$ is the vocabulary of the feature $j$ (unique categories)
$$P(x_j = v|y_k) = \frac{\#(y = k \wedge x_j = v) + \alpha}{\#(y = k) + \alpha|V_j|}$$
- sklearn.naive_bayes.CategoricalNB

# Naïve Bayes: Theoretical formulation

## Likelihood: Multinomial distribution (text)

$$p(word = w_j | y = k) = \frac{C_{jk} + \alpha}{C_k + \alpha |V|}$$

$$\hat{y} = \arg\max_k p(y = k) * \prod_{w \in text} p(word = w | y = k)$$

- A text can be seen as one feature with words as values
- $C_k$ is the number of training samples having $k$ as class
- $C_{jk}$ is the number of occurrences of word $w_j$ in texts having $k$ as class
- $V$ is the vocabulary (unique words in the training dataset)
- sklearn.naive_bayes.MultinomialNB

# Naïve Bayes: Theoretical formulation

## Likelihood: Bernoulli distribution

$$p(x_j = v|y = k) = p(x_j = 1|y = k)v + (1 - p(x_j = 1|y = k))(1 - v)$$

$$p(x_j = 1|y_k) = \frac{|\{x_j^{(i)} = 1 \wedge y^{(i)} = k,\ i \in \{1, \cdots, M\}\}|}{|\{y^{(i)} = k,\ i \in \{1, \cdots, M\}\}|}$$

- $x_j$ is a boolean feature having a value $v \in \{0, 1\}$
- sklearn.naive_bayes.BernoulliNB

# Naïve Bayes: Theoretical formulation

## Likelihood: Normal (Gaussian) distribution

$$p(x_j = v|Y_k) = \frac{1}{\sqrt{2\pi\sigma_{kj}^2}} e^{\frac{-(v-\mu_{kj})^2}{2\sigma_{kj}^2}}$$

- $x_j$ is a numerical feature having values $v \in ]-\infty, +\infty[$
- $\mu_{kj}$ is the mean of $x_j$'s values having $k$ as class
- $\sigma_{kj}^2$ is the **unbiased** variance of $x_j$'s values having $k$ as class
- sklearn.naive_bayes.GaussianNB

This slide
is left blank
for no reason

Section 2

# Numerical application

# Naïve Bayes
## Numerical application

- **Multinomial NB**: Play or not based on these categorical features: outlook (sunny, overcast, rainy), temp (hot, mild, cool), humidity (high, normal), windy (true, false)
- **Bernoulli NB**: Pass the exam or fail based on these boolean features: confident, studied, sick
- **Normal NB**: Male or female based on these numerical features: height (cm), weight (kg), footsize (cm)

# Naïve Bayes: Numerical application

## Multinomial NB: Example (1)

| outlook | temp | humidity | windy | play |
|---|---|---|---|---|
| sunny | hot | high | false | no |
| sunny | hot | high | true | no |
| overcast | hot | high | false | yes |
| rainy | mild | high | false | yes |
| rainy | cool | normal | false | yes |
| rainy | cool | normal | true | no |
| overcast | cool | normal | true | yes |
| sunny | mild | high | false | no |
| sunny | cool | normal | false | yes |
| rainy | mild | normal | false | yes |
| sunny | mild | normal | true | yes |
| overcast | mild | high | true | yes |
| overcast | hot | normal | false | yes |
| rainy | mild | high | true | no |

- Prior probability
  - $p(play = yes) = \frac{\#(play=yes)}{M} = \frac{9}{14}$
  - $p(play = no) = \frac{\#(play=no)}{M} = \frac{5}{14}$
- Likelihood probability of $outlook = rainy$
  - $p(outlook = rainy|play = yes) = \frac{\#(outlook=rainy \wedge play=yes)}{\#(play=yes)} = \frac{3}{9}$
  - $p(outlook = rainy|play = no) = \frac{\#(outlook=rainy \wedge play=no)}{\#(play=yes)} = \frac{2}{5}$
- Likelihood probability of $temp = hot$
  - $p(temp = hot|play = yes) = \frac{\#(temp=hot \wedge play=yes)}{\#(play=yes)} = \frac{2}{9}$
  - $p(temp = hot|play = no) = \frac{\#(temp=hot \wedge play=no)}{\#(play=yes)} = \frac{2}{5}$

# Naïve Bayes: Numerical application
## Multinomial NB: Example (2)

- Likelihood probability of $humidity = high$
  - $p(humidity = high|play = yes) = \frac{\#(humidity=high \wedge play=yes)}{\#(play=yes)} = \frac{3}{9}$
  - $p(humidity = high|play = no) = \frac{\#(humidity=high \wedge play=no)}{\#(play=yes)} = \frac{4}{5}$
- Likelihood probability of $windy = false$
  - $p(windy = false|play = yes) = \frac{\#(windy=false \wedge play=yes)}{\#(play=yes)} = \frac{6}{9}$
  - $p(windy = false|play = no) = \frac{\#(windy=false \wedge play=no)}{\#(play=yes)} = \frac{2}{5}$

Given $\vec{v} = [rainy, hot, high, false]$

- $p(play = yes|x = \vec{v}) \propto \frac{9}{14}(\frac{3}{9}\frac{2}{9}\frac{3}{9}\frac{6}{9}) = \frac{6}{567} \approx 0.0106$
- $p(play = no|x = \vec{v}) \propto \frac{5}{14}(\frac{2}{5}\frac{2}{5}\frac{4}{5}\frac{2}{5}) = \frac{16}{875} \approx 0.0183$
- $\hat{y} = no$

# Naïve Bayes: Numerical application

## Bernoulli NB: Example (1)

| confident | studied | sick | result |
|---|---|---|---|
| 1 | 0 | 0 | fail |
| 1 | 0 | 1 | pass |
| 0 | 1 | 1 | fail |
| 0 | 1 | 0 | pass |
| 1 | 1 | 1 | pass |

- Prior probability
  - $p(pass) = \frac{\#(pass)}{M} = \frac{3}{5}$
  - $p(fail) = \frac{\#(fail)}{M} = \frac{2}{5}$

- Prior probability
  - $p(confident|pass) = \frac{2}{3}$
  - $p(studied|pass) = \frac{2}{3}$
  - $p(sick|pass) = \frac{2}{3}$
  - $p(confident|fail) = \frac{1}{2}$
  - $p(studied|fail) = \frac{1}{2}$
  - $p(sick|fail) = \frac{1}{2}$

Given $\vec{v} = [1, 0, 0]$

- $p(pass|\vec{v}) \propto \frac{3}{5}[\frac{2}{3}(1 - \frac{2}{3})(1 - \frac{2}{3})] = \frac{2}{45} \approx 0.0444$
- $p(fail|\vec{v}) \propto \frac{2}{5}[\frac{1}{2}(1 - \frac{1}{2})(1 - \frac{1}{2})] = \frac{1}{20} \approx 0.05$
- $\hat{y} = fail$

# Naïve Bayes: Numerical application

## Normal NB: Example

| height | weight | footsize | person |
|---|---|---|---|
| 182 | 81.6 | 30 | male |
| 180 | 86.2 | 28 | male |
| 170 | 77.1 | 30 | male |
| 180 | 74.8 | 25 | male |
| 152 | 45.4 | 15 | female |
| 168 | 68.0 | 20 | female |
| 165 | 59.0 | 18 | female |
| 175 | 68.0 | 23 | female |

- Prior probability: no need since the classes distribution is uniform

| person | height | | weight | | footsize | |
|---|---|---|---|---|---|---|
| | $\mu$ | $\sigma^2$ | $\mu$ | $\sigma^2$ | $\mu$ | $\sigma^2$ |
| male | 178 | 29.33 | 79.92 | 25.48 | 28.25 | 5.58 |
| female | 165 | 92.67 | 60.1 | 114.04 | 19 | 11.33 |

Given $\vec{v} = [183, 59, 20]$

- $p(height = 183|male) = \frac{1}{\sqrt{2\pi * 29.33}} e^{\frac{-(183-178)^2}{2*29.33}} \approx 0.04810173$
- $p(height = 183|female) = \frac{1}{\sqrt{2\pi * 92.67}} e^{\frac{-(183-165)^2}{2*92.67}} \approx 0.00721463$

Section 3

# Bibliography

# Bibliography

Metsis, V., Androutsopoulos, I., and Paliouras, G. (2006).
Spam filtering with naive bayes - which naive bayes?
In *International Conference on Email and Anti-Spam*.

The probability of finding
another slide
given
the slides you already seen
is
0