Lecture 6: Naive Bayes classifier

iie Robert

classifiers

Bayes classifier

framework

Example: Dog

Hyperparameters

Advantage and limits

Further algorithms

# Lecture 6: Naive Bayes classifier Introduction to Machine Learning

Sophie Robert

L3 MIASHS — Semestre 2

2022-2023

#### Lecture 6: Naive Bayes classifier

Sophie Robert

Principle of

Principle of Bayes classifier

Mathematical

Example: Dog

Hyperparameters

Advantages and limits

Further algorithms

- 1 Probabilistic classifiers
- 2 Principle of Bayes classifier
- 3 Mathematical framework
- 4 Example: Dog breed prediction
- 5 Hyperparameters
- 6 Advantages and limits
- 7 Further algorithms

Lecture 6: Naive Bayes classifier

oop...e mobe

# Probabilistic classifiers

Principle of Bayes classifie

Mathematical

Example: Dog breed

Hyperparameters

Advantage and limits

Further

#### Probabilistic classifiers

Probabilistic classifiers are classifiers that predict, given an observation of an input, a **probability distribution over a set of classes** (instead of simply the class like standard classifiers).

Lecture 6: Naive Bayes classifier

Sopille Rober

# Probabilistic classifiers

Principle of Bayes classifier

Mathematica

Example: Dog

breed prediction

Hyperparameters

Advantages and limits

Further

#### Probabilistic classifiers

Probabilistic classifiers are classifiers that predict, given an observation of an input, a **probability distribution over a set of classes** (instead of simply the class like standard classifiers).

Given a record  $\mathbf{x} = (x_1, x_2, ..., x_n) \in \mathbb{R}^n$  and a set of labels  $y_i \in \mathcal{Y}$ , provide an estimation of  $\mathbb{P}(y_i|\mathbf{x})$   $y_i \in \mathcal{Y}$  (and assign most likely label to  $\mathbf{x}$ ).

Lecture 6: Naive Bayes classifier

hie Robert

Probabilistic classifiers

Principle of Bayes classifier

framework

Example: Dog breed

Hyperparameters

Advantage and limits

Further

#### Question

What is in your opinion one of the strength of working with a probabilistic model rather than a classification function?

Lecture 6: Naive Bayes classifier

Sophie Rober

Probabilistic classifiers

Principle of Bayes classifier

Mathematical framework

Example: Dog breed prediction

Hyperparameters

Advantages

Further

#### Question

What is in your opinion one of the strength of working with a probabilistic model rather than a classification function?

Possible native models are:

- Logistic regression
- Subtypes of neural networks
- Bayes classifiers

Lecture 6: Naive Bayes classifier

#### Probabilistic classifiers

Hyperparameters

#### Question

What is in your opinion one of the strength of working with a probabilistic model rather than a classification function?

Possible native models are:

- Logistic regression
- Subtypes of neural networks
- Bayes classifiers

Non-probabilistic models can also be turned into a probabilistic classifier (SVM, trees . . . ).

#### Main idea

Lecture 6: Naive Bayes classifier

hie Robert

Principle of

Bayes classifier

Mathematical

Example: Dog breed

Hyperparameters

Advantage

Further

#### The naïve Bayes classifier algorithm

The naïve Bayes classifier algorithm is a a probabilistic classifier based on applying Bayes' theorem with independence assumptions between the features.

#### Main idea

Lecture 6: Naive Bayes classifier

nie Robert

Principle of

Bayes classifier

framework

breed prediction

Hyperparameters

Advantages

Further

#### The naïve Bayes classifier algorithm

The naïve Bayes classifier algorithm is a a probabilistic classifier based on applying Bayes' theorem with independence assumptions between the features.

Each features **contribute to the class probability independently**: for example, the size and the weight of the dog contribute independently to its breed.

Lecture 6: Naive Bayes classifier

nhie Robert

Probabilisti classifiers

Principle of Bayes classifier

Mathematical framework

Example: Dog breed

Hyperparameters

Advantage

Further

We want to estimate for each label  $y_i \in \mathcal{Y} \mathbb{P}(y_i|\mathbf{x})$  (what is the probability of being label  $y_i$  given the data records ?).

Lecture 6: Naive Bayes classifier

phie Robert

classifiers

Principle of Bayes classifier

Mathematical framework

Example: Do breed

Hyperparameters

Advantage

Further

We want to estimate for each label  $y_i \in \mathcal{Y} \mathbb{P}(y_i|\mathbf{x})$  (what is the probability of being label  $y_i$  given the data records ?) . However, if n is large, the computation is infeasible.

Lecture 6: Naive Bayes classifier

ophie Rober

classifiers

Bayes classifie

Mathematical framework

Example: Dog breed

Hyperparameters

Advantages and limits

Further algorithms We want to estimate for each label  $y_i \in \mathcal{Y} \mathbb{P}(y_i|\mathbf{x})$  (what is the probability of being label  $y_i$  given the data records?) . However, if n is large, the computation is infeasible.

Using the definition of conditional probabilities:

$$\mathbb{P}(y_i|\mathbf{x}) = \frac{\mathbb{P}(y_i,\mathbf{x})}{\mathbb{P}(\mathbf{x})}$$

 $P(\mathbf{x})$  is a constant because  $\mathbf{x}$  is given so we only need to find the value of  $\mathbb{P}(y_i, \mathbf{x})$ .

Lecture 6: Naive Bayes classifier

Sophie Robert

Probabilisti

Principle of

Mathematical

framework

Example: Dog breed

Hyperparameters

Advantage and limits

Further

Using the definition of conditional probabilities iteratively,

Lecture 6: Naive Bayes classifier

hie Robert

classifiers

Principle of Bayes classifier

framework

Example: Dog
breed

Hyperparameters

Advantages and limits

Further algorithms

Using the definition of conditional probabilities iteratively,

$$\mathbb{P}(y_i, \mathbf{x}) = \mathbb{P}(x_1, x_2, ..., x_n, y_i) 
= \mathbb{P}(x_1 | x_2, x_3, ..., y_i) \times \mathbb{P}(x_2, x_4, ..., y_i) 
= \mathbb{P}(x_1 | x_2, x_3, ..., y_i) \times \mathbb{P}(x_2 | x_3, x_4, ..., y_i) \times \mathbb{P}(x_3, x_4, ..., y_i) 
= ... 
= \mathbb{P}(x_1 | x_2, x_3, ..., y_i) \times \mathbb{P}(x_2 | x_3, x_4, ..., y_i) \times \mathbb{P}(x_n | y_i) \times \mathbb{P}(y_i)$$

Lecture 6: Naive Bayes classifier

Sophie Robert

Classitiers
Principle of

Bayes classifie

Mathematical framework

Example: Dog breed prediction

Hyperparameters

Advantages and limits

Further algorithm We now make the hypothesis that each feature  $x_i$  are **conditionnally independant** given  $y_i$  (and only depends on the label  $y_i$ ):

#### Conditional independence

**Conditional independence** describes situations where an observation is irrelevant: the probability of the hypothesis given the uninformative observation is equal to the probability without. If A is the hypothesis, B and C the observations,  $P(A \mid B, C) = P(A \mid C)$ 

Lecture 6: Naive Bayes classifier

Sophie Robert

Principle of

Bayes classifie

framework

Example: Dog

breed prediction

Hyperparameters

Advantages and limits

Further algorithm We now make the hypothesis that each feature  $x_i$  are **conditionnally independant** given  $y_i$  (and only depends on the label  $y_i$ ):

#### Conditional independence

**Conditional independence** describes situations where an observation is irrelevant: the probability of the hypothesis given the uninformative observation is equal to the probability without. If A is the hypothesis, B and C the observations,  $P(A \mid B, C) = P(A \mid C)$ 

$$\mathbb{P}(x_1|x_2, x_3, ..., y_i) = \mathbb{P}(x_1|y_i)$$

Lecture 6: Naive Bayes classifier

Sophie Robert

Probabilisti classifiers

Principle of Bayes classifier

Mathematical framework

breed prediction

Hyperparameters

and limits

Further algorithms

We now have:

$$\mathbb{P}(y_i, \mathbf{x}) = \mathbb{P}(y_i) \prod_{j=1}^n \mathbb{P}(x_j | y_i)$$

and:

$$\mathbb{P}(y_i|\mathbf{x}) \propto \mathbb{P}(y_i) \prod_{i=1}^n \mathbb{P}(x_i|y_i)$$

Lecture 6: Naive Bayes classifier

Sophie Robert

Probabilisti classifiers

Principle of Bayes classifier

Mathematical framework

Example: Dog breed

Hyperparameters

Advantages and limits

Further algorithms

We now have:

$$\mathbb{P}(y_i, \mathbf{x}) = \mathbb{P}(y_i) \prod_{j=1}^n \mathbb{P}(x_j | y_i)$$

and:

$$\mathbb{P}(y_i|\mathbf{x}) \propto \mathbb{P}(y_i) \prod_{i=1}^n \mathbb{P}(x_i|y_i)$$

We then select the most probable class

$$\hat{y} = argmax_{i=1,...,k}(\mathbb{P}(y_i) \prod_{i=1}^{n} \mathbb{P}(x_j|y_i))$$

Lecture 6: Naive Bayes classifier

hie Robert

Probabilisti classifiers

Principle of

Mathematical

framework

Example: Dog breed

Hyperparameters

Advantages

Further

Can you guess why this algorithm can be called naive ?

Lecture 6: Naive Bayes classifier

Probabilistic

Principle of Bayes classifie

Mathematical framework

Example: Dog

Hyperparameters

Advantages

Further algorithm

Can you guess why this algorithm can be called naive?

We have two terms to estimate:

- $\mathbb{P}(y_i)$ : either assume class equiprobability or estimate using the frequency in training dataset
- $\mathbb{P}(x_j|y_i)$ : we need to decide on a conditional law

Lecture 6: Naive Bayes classifier

Sophie Robert

Probabilisti classifiers

Bayes classifie

Mathematical framework

Example: Do breed

Hyperparameters

Advantage and limits

Further algorithm

Possible assumptions include:

■ If  $X_j$  is a continuous variable  $(\mathbf{x_j} \in \mathbb{R})$ , the continuous values associated within class i are distributed according to a Gaussian distribution parametrized with mean  $\mu_i$  and variance  $\sigma_i$ 

$$f(v \mid y_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(v-\mu_i)^2}{2\sigma_i^2}}$$

Lecture 6: Naive Bayes classifier

Sophie Robert

Probabilisti classifiers

Principle of Bayes classifier

Mathematical framework

Example: Dog breed

Hyperparameters

Advantages

Further algorithm Possible assumptions include:

■ If  $X_j$  is a continuous variable  $(\mathbf{x_j} \in \mathbb{R})$ , the continuous values associated within class i are distributed according to a Gaussian distribution parametrized with mean  $\mu_i$  and variance  $\sigma_i$ 

$$f(v \mid y_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(v-\mu_i)^2}{2\sigma_i^2}}$$

■ If  $X_j$  is a categorial variable  $(x_j \in \{0, ..., K\})$ , the probability can be estimated as the proportion of values within class:

$$\mathbb{P}(x=j|y_i)=\frac{N_{ji}}{N_i}$$

# Example

Lecture 6: Naive Bayes classifier

Sophie Rober

Probabilistic classifiers

Principle of Bayes classifie

Bayes classifie
Mathematical

Example: Dog breed prediction

Hyperparameters

Advantages

Further algorithms

#### Training dataset:

| Training databets |        |      |                |  |  |  |
|-------------------|--------|------|----------------|--|--|--|
| Height            | Weight | Tail | Label          |  |  |  |
| 45                | 30     | 0    | Labradoodle    |  |  |  |
| 30                | 25     | 1    | Labradoodle    |  |  |  |
| 40                | 35     | 1    | Labradoodle    |  |  |  |
| 20                | 15     | 0    | English cocker |  |  |  |
| 22                | 18     | 1    | English cocker |  |  |  |
| 25                | 20     | 1    | English cocker |  |  |  |
|                   |        |      |                |  |  |  |

#### Individual to classify

| Height | Weight | Tail | Label |
|--------|--------|------|-------|
| 25     | 31     | 1    | ?     |

# Example: training the model

Lecture 6: Naive Bayes classifier

Sophie Robert

classifiers

Principle of Bayes classifie

Example: Dog

prediction

Hyperparameters

A. . .

Further

#### Training the model

Training the model consists in computing **statistical estimators over the population**.

For the labradoodle population:

|      | Height | Weight |
|------|--------|--------|
| Mean | 38.33  | 30     |
| Var  | 38.89  | 16.67  |

For the cocker population:

| Tor the cocker population. |        |        |  |  |  |
|----------------------------|--------|--------|--|--|--|
|                            | Height | Weight |  |  |  |
| Mean                       | 22.33  | 17.66  |  |  |  |
| Var                        | 4.22   | 4.22   |  |  |  |

# Example: solution

Lecture 6: Naive Bayes classifier

Sophie Rober

Probabilistic classifiers

Principle of Bayes classifier

Mathematical framework

Example: Dog breed prediction

Hyperparameters

Advantages

Further algorithms

#### **Estimate:**

$$\begin{split} &\mathbb{P}(\mathsf{labradoodle} \mid \mathsf{height} = 25, \mathsf{weight} = 31, \mathsf{tail} = 1) \\ &\propto \mathbb{P}(\mathsf{labradoodle}) \times \mathbb{P}(\mathsf{height} = 25 | \mathsf{labradoodle}) \\ &\times \mathbb{P}(\mathsf{weight} = 31 | \mathsf{labradoodle}) \\ &\times \mathbb{P}(\mathsf{tail} = 1 | \mathsf{labradoodle}) \end{split}$$

$$\mathbb{P}(\text{labradoodle}) = \frac{1}{2}$$

$$\mathbb{P}(\text{height} = 25 | \text{labradoodle}) = \frac{1}{2}$$

$$\mathbb{P}(\text{height} = 25|\text{labradoodle}) = \frac{1}{\sqrt{2\pi \times 38.89}} e^{-\frac{(25 - 38.33)^2}{2 \times 38.89}} = 0.006$$

$$\mathbb{P}(\text{weight} = 31 | \text{labradoodle}) = \frac{1}{\sqrt{2\pi \times 16.67}} e^{-\frac{(31-30)^2}{2\times 16.67}} = 0.09$$

$$\mathbb{P}(\mathsf{tail} = 1 | \mathsf{labradoodle}) = \frac{2}{3}$$

$$\mathbb{P}(\mathsf{labradoodle} \mid \mathsf{height} = 25, \mathsf{weight} = 31, \mathsf{tail} = 1) = 0.00017$$

# Example: solution

Lecture 6: Naive Bayes classifier

Sophie Rober

Probabilisti classifiers

Principle of Bayes classifier

Mathematical

Example: Dog breed prediction

Hyperparameters

Advantages and limits

Further algorithms

#### Estimate:

$$\begin{split} &\mathbb{P}(\mathsf{cocker} \mid \mathsf{height} = 25, \mathsf{weight} = 31, \mathsf{tail} = 1) \\ &\propto \mathbb{P}(\mathsf{cocker}) \times \mathbb{P}(\mathsf{height} = 25 | \mathsf{cocker}) \\ &\times \mathbb{P}(\mathsf{weight} = 31 | \mathsf{cocker}) \\ &\times \mathbb{P}(\mathsf{tail} = 1 | \mathsf{cocker}) \end{split}$$

$$\mathbb{P}(\mathsf{cocker}) = \frac{1}{2}$$

$$\mathbb{P}(\mathsf{height} = 25 | \mathsf{cocker}) = \frac{1}{\sqrt{2\pi \times 4.22}} e^{-\frac{(25 - 22.33)^2}{2\times 4.22}} = 0.08$$

$$\mathbb{P}(\text{weight} = 31|\text{cocker}) = \frac{1}{\sqrt{2\pi \times 4.22}} e^{-\frac{(31-17.66)^2}{2\times 4.22}} = 1.31e - 11$$

$$\mathbb{P}(\mathsf{tail} = 1 | \mathsf{cocker}) = \frac{2}{3}$$

$$\mathbb{P}(\text{cocker} \mid \text{height} = 25, \text{weight} = 31, \text{tail} = 1) = 0.00$$

# Hyperparameters

Lecture 6: Naive Bayes classifier

Sophie Rober

Probabilist classifiers

Principle of Bayes classifier

Mathematical

Example: Dog breed

prediction

Hyperparameter

Advantages and limits

Further algorithms

#### Hyperparameters

What **hyperparameters\*** do the naive Bayes classifier require?

Lecture 6: Naive Bayes classifier

Sophie Robert

Probabilist classifiers

Principle of Bayes classifier

Mathematica framework

Example: Dog breed

Hyperparameters

Advantages and limits

Further

#### Limits:

Lecture 6: Naive Bayes classifier

Sophie Robert

classifiers

Principle of Bayes classifie

Example: Dog

prediction

Hyperparameters

Advantages and limits

Further

#### Limits:

- Strong independance hypothesis (but in practice, naive bayes behave rather well)
- Unable to classify unknown classes that do not show in training set (always sets it to 0, except in the case of artificial equiprobability)

Lecture 6: Naive Bayes classifier

Sophie Robert

classifiers

Principle of Bayes classifie

Example: Dog

Hyperparameters

Advantages and limits

Further algorithms

#### Limits:

- Strong independance hypothesis (but in practice, naive bayes behave rather well)
- Unable to classify unknown classes that do not show in training set (always sets it to 0, except in the case of artificial equiprobability)

#### Advantages:

Lecture 6: Naive Bayes classifier

D 1 1 1 1 1 1 1 1

Principle of Bayes classifie

Bayes classifie

Example: Dog breed

Hyperparameters

Advantages and limits

Further algorithms

#### Limits:

- Strong independance hypothesis (but in practice, naive bayes behave rather well)
- Unable to classify unknown classes that do not show in training set (always sets it to 0, except in the case of artificial equiprobability)

#### Advantages:

- Extends naturally to multi-class
- Naturally deals with categorical variables

# Other probabilistic classifiers

Lecture 6: Naive Bayes classifier

ophie Rober

Probabilistic classifiers

Principle of Baves classifie

Mathematical

Example: Dog breed

Hyperparameters

Advantage and limits

Further algorithms

One of the most famous probabilistic classifier is **logistic regression**: probability distribution is expressed as the logit of the linear combination of features.

# Other probabilistic classifiers

Lecture 6: Naive Bayes classifier

Hyperparameters

Further algorithms One of the most famous probabilistic classifier is **logistic** regression: probability distribution is expressed as the logit of the linear combination of features.

Using **softmax function** as activation layer in **neural** networks transforms output into a probability distribution consisting of K probabilities proportional to the exponentials of the input numbers.

# Questions

Lecture 6: Naive Bayes classifier

Sophie Robert

Probabilist classifiers

Principle of Bayes classifi

Mathematica framework

Example: Dog

Hyperparameters

Advantage

Further algorithms

## Questions?