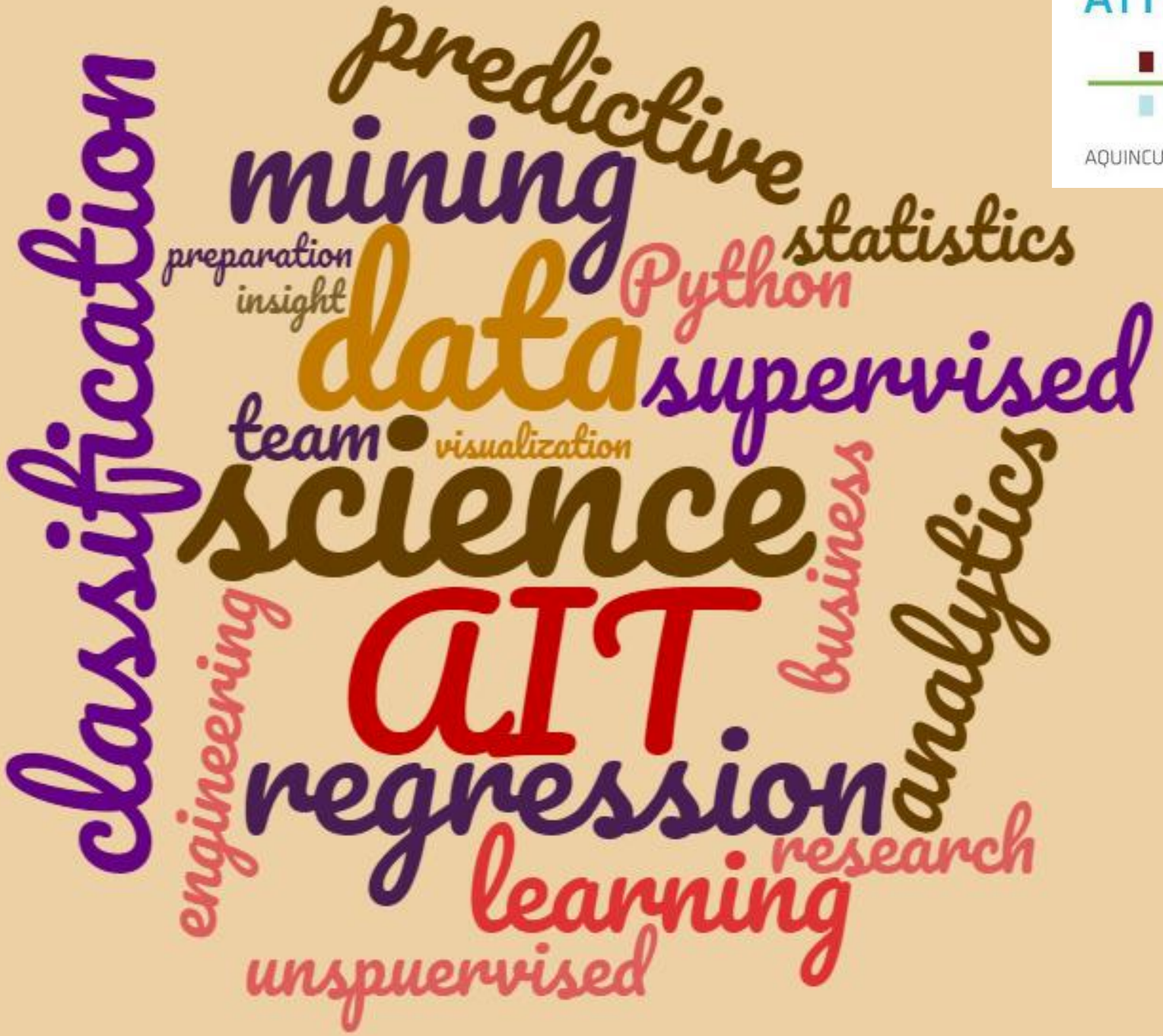
Data Science

March 10, 2020.
Naive Bayes



AIT-BUDAPEST
AQUINCUM INSTITUTE OF TECHNOLOGY

Roland Molontay

Principle of Bayes classifier

- The attributes and class label are considered as random variables
  - Let C denote the class label and $A_1, A_2, \ldots, A_d$ denote the attributes
  - Earlier we use Y for the target variable (label) and X_i for the attributes
    - First, for the sake of simplicity we consider the attributes as categorical variables
- We learnt that the misclassification error is minimal if we choose the class label to maximize $P(C \mid A_1, A_2, \ldots, A_d)$ conditional probability
  - $c^* = \underset{c}{\operatorname{argmax}} \mathrm{P}(\mathrm{C} = \mathrm{c} \mid \underline{A} = \underline{a})$
  - Maximum *a posteriori estimation*
- Can we estimate the value of $P(C \mid A_1, A_2, \ldots, A_d)$ from the data?

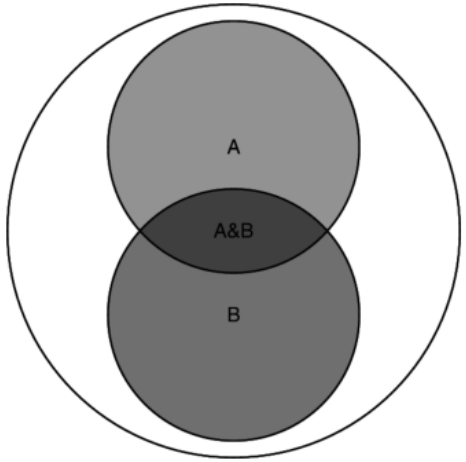
Quick revision of necessary notions

- Conditional probability

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

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

$$P(A \mid B^c) = \frac{P(A,B^c)}{P(B^c)}$$



- Law of total probability ($\{B_i : i = 1,2,3, \ldots\}$ is a partition of the sample space)

$$P(A) = \sum_i P(A, B_i) = \sum_i P(B_i)P(A|B_i)$$

- Bayes theorem

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)} = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid B^c)P(B^c)}$$

Using Bayes theorem for classification

- Use Bayes theorem:

$$c^* = \underset{c}{\operatorname{argmax}}\, \mathrm{P}(\mathrm{C} = \mathrm{c} \mid \underline{A} = \underline{a}) = \underset{c}{\operatorname{argmax}} \frac{\mathrm{P}(\underline{A} = \underline{a} \mid C = c) \cdot P(C = c)}{\mathrm{P}(\underline{A} = \underline{a})} =$$

$$= \underset{c}{\operatorname{argmax}}\, \mathrm{P}(\underline{A} = \underline{a} \mid C = c) \cdot P(C = c)$$

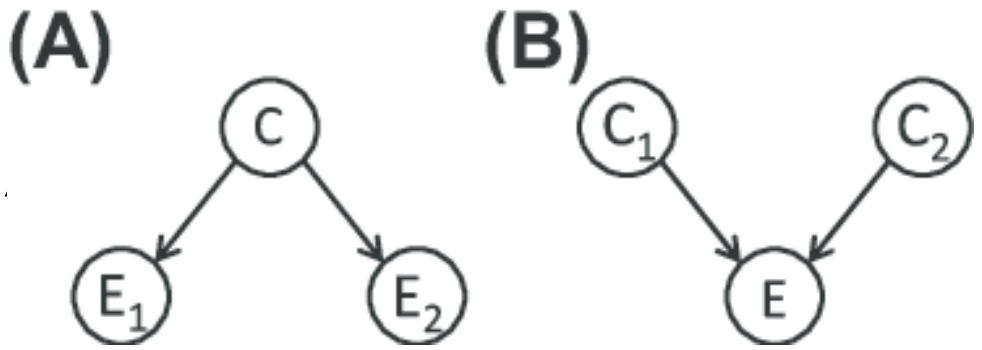
  - The first identity uses Bayes theorem. The second identity is valid because the denominator does not depend on the class label, it is the same for any C values
- If $P(C = c)$ is constant for any class label, i.e. the *a priori* probabilites agree for all labels, then
  $c^* = \underset{c}{\operatorname{argmax}}\, \mathrm{P}(\mathrm{C} = \mathrm{c} \mid \underline{A} = \underline{a}) = \underset{c}{\operatorname{argmax}}\, \mathrm{P}(\underline{A} = \underline{a} \mid C = c)$, thus maximum *a posteriori* (the first) and maximum likelihood (the second) estimations are equal

Estimating probabilities from data

- Based on Bayes theorem: $c^* = \underset{c}{\operatorname{argmax}} \mathrm{P}\left(\underline{A} = \underline{a} \,\middle|\, C = c\right) \cdot P(C = c)$
- How do we estimate the probabilities?
- $P\left(C = c_j\right) = \frac{n_j}{n}$, i.e. the relative frequency of records with label c_j to the number of total records
- To estimate $P\ (A_1, A_2, \dots, A_d \mid C)$ we have to estimate a lot of parameters
  - For binary attributes: $\#c \cdot (2^d - 1)$ parameters
  - It is only possible if we have very big data
- Naive Bayes (Naïve Bayes) approach:
  we assume that $A_1, A_2, \dots, A_d$ random variables are conditionally independent of each other given the class variable C
  $P\ (A_1, A_2, \dots, A_n \mid C) = \mathrm{P}(\mathrm{A}_1 \mid \mathrm{C}) \cdot\ \mathrm{P}(\mathrm{A}_2 \mid \mathrm{C}) \cdot \dots \cdot \mathrm{P}(\mathrm{A}_\mathrm{d} \mid \mathrm{C})$
  - We have to estimate $P(A_i = a_i \mid C = c_j)$ values for any possible *i, j* pairs
  - Using Naive Bayes assumption the complexity of the problem is reduced
    - For binary attributes there are $\#c \cdot d$ parameters to fit

Naive Bayes assumption

- Naive Bayes assumption: the features are conditionally independent of each other given the class variable
- Conditionally independent but not independent variables:
  - The reading ability of a child is conditionally independent of his/her height given his/her age (common cause – Fig. A)
  - Thunder is independent of rain, given lightning
- Independent but not conditionally independent variables:
  - Basketball game is independent of rain but they are not conditionally independent given the traffic (common effect – Fig. B)



Estimating probabilities from data II.

- Calculating (estimating) $P(A_i = a_i | C = c_j)$ for categorical attributes $P(A_i = a_i | C = c_j) = \frac{n_{ij}}{n_j}$, i.e. number of records with label c_j and with *i*th attribute a_i divided by the number of records with label c_j
- For continuous attributes
  - Transforming it to discrete
  - Distribution based estimation
    - Approximation by normal distribution

Approximation by normal distribution

- We assume that

$$P(A_i = a_i \mid C = c_j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \mathrm{e}^{-\frac{(a_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

  - We have to estimate the value of σ_{ij} and μ_{ij} from the data
    - Sample mean and sample standard deviation
  - μ_{ij} is the mean of attribute A_i considering the records with class c_j
  - σ_{ij} is the standard deviation of attribute A_i among records with class c_j
- Be careful: the formula above is quite sloppy from a mathematical point of view, regarding continuous variables the question itself is meaningless (instead of a concrete a_i value it would be meaningful to ask for an interval)
  - However, in practice it is used that way!

Example for probability estimation

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Estimating the probability of class

  P(No) = 7/10,
  P(Yes) = 3/10

- Estimating the probability of attribute values (discrete case)

  P(Status=Married|No) = 4/7
  P(Refund=Yes|Yes)=0

Example for probability estimation II.

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Estimating the probability of attribute values (continuous case, normal approximation)
  - If Class = No than for Income
    - Mean = 110
    - Sample variance = 2975

$$P(Income = 120 \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

Prediction

- After we estimated the conditional probabilities based on the training set then we can classify a new observation based on its $(a_1, a_2, \dots, a_d)$ feature vector
- For every c_j label we calculate the following probability:

$$\mathrm{P}\left(\underline{A} = (a_1, a_2, \dots, a_d) \middle| C = c_j\right) \cdot P(C = c_j) =$$
$$\mathrm{P}(\mathrm{A}_1 = \mathrm{a}_1 \,|\mathrm{C} = \mathrm{c_j}) \cdot \ \mathrm{P}(\mathrm{A}_2 = a_2 \,|\mathrm{C} = \mathrm{c_j}) \cdot \dots \cdot \mathrm{P}(\mathrm{A_d} = a_d \,|\mathrm{C} = \mathrm{c_j}) \cdot P(C = c_j)$$

  - We predict the c_j class label for which the probability above is maximal, that was our aim to find the maximum *a posteriori* estimation:

$$c^* = \underset{c}{\operatorname{argmax}}\, \mathrm{P}\left(\underline{A} = \underline{a} \middle| C = c\right) \cdot P(C = c)$$

Example for classification

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

A: attributes

M: mammals

N: non-mammals

$$P(A \mid M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A \mid N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A \mid M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A \mid N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

P(A|M)P(M) > P(A|N)P(N)

=> Mammals

Example for classification II.

$$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 120\text{K})$$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

naive Bayes Classifier:

P(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
P(Refund=Yes|Yes) = 0
P(Refund=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/7
P(Marital Status=Divorced|Yes)=1/7
P(Marital Status=Married|Yes) = 0

For taxable income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

- P(X|Class=No) = P(Refund=No|Class=No)
  × P(Married| Class=No)
  × P(Income=120K| Class=No)
  = 4/7 × 4/7 × 0.0072 = 0.0024

- P(X|Class=Yes) = P(Refund=No| Class=Yes)
  × P(Married| Class=Yes)
  × P(Income=120K| Class=Yes)
  = 1 × 0 × 1.2 × 10^{-9} = 0

Since P(X|No)P(No) > P(X|Yes)P(Yes)

Therefore P(No|X) > P(Yes|X)

=> Class = No

What happens if a conditional probability is 0?

- It may occur that for an *i, j* pair the estimated probability is $\mathrm{P}(\mathrm{A_i} = \mathrm{a_i} \,|\mathrm{C} = \mathrm{c_j}) = 0$, since there are no such records in the training set
  - Based on the other attributes a certain class may look a good choice but if a conditional probability is estimated to be 0 that makes the whole expression to be 0
  - Solution: using different estimation methods with smoothing!
- Original: $P(A_i = a_i | \, C = c_j) = \frac{n_{ij}}{n_j}$
- Laplace smoothing: $P(A_i = a_i | \, C = c_j) = \frac{n_{ij} + 1}{n_j + \#c}$
- m-smoothing: $P(A_i = a_i | \, C = c_j) = \frac{n_{ij} + mp}{n_j + m}$
  - Where: n_{ij}: number of records with label c_j and with *i*th attribute a_i; n_j: the number of records with label c_j; #c: number of possible labels, p: prior expectation for probability, m: a parameter

Example: classifying mammals with Laplace smoothing

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
eagle	no	yes	no	yes	non-mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

$$P(A \mid M) = \frac{4}{5} \times \frac{4}{5} \times \frac{0}{5} \times \frac{0}{5} = 0$$

$$P(A \mid N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A \mid M)P(M) = 0$$

$$P(A \mid N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

Without Laplace,

P(A|M)P(M) < P(A|N)P(N)

=> Non-Mammals

$$P(A \mid M) = \frac{5}{6} \times \frac{5}{6} \times \frac{1}{6} \times \frac{1}{6} = 0.019$$

$$P(A \mid N) = \frac{2}{14} \times \frac{11}{14} \times \frac{4}{14} \times \frac{5}{14} = 0.011$$

$$P(A \mid M)P(M) = 0.019 \times \frac{8}{22} = 0.0070$$

$$P(A \mid N)P(N) = 0.011 \times \frac{14}{22} = 0.0072$$

With Laplace,

P(A|M)P(M) ~ P(A|N)P(N)

=> Cannot really decide

In this example the denominator is increased by 1, but according to the formula it should have been increased by the number of labels (i.e. by 2)!

Evaluation of Naive Bayes

- Robust (insensitive) for irrelevant attributes
- It can handle missing data (it ignores the missing values for the probability estimation)
- The naive Bayes assumption (conditional independence) in many cases is not valid in reality
  - In practice it still works quite well
  - This assumption can be refined (Bayesian Networks, Bayesian Belief Networks)

Problem

Classify the following record X = (Marital status = Single, Annual income = 90K) using the Naive Bayes classifier based on the training data in the following table, where Default is the class label. Discretize annual income by 20K intervals: [60K, 80K), [80K, 100K), . . . !

1. Use the original estimates!
2. Use Laplace smoothing!

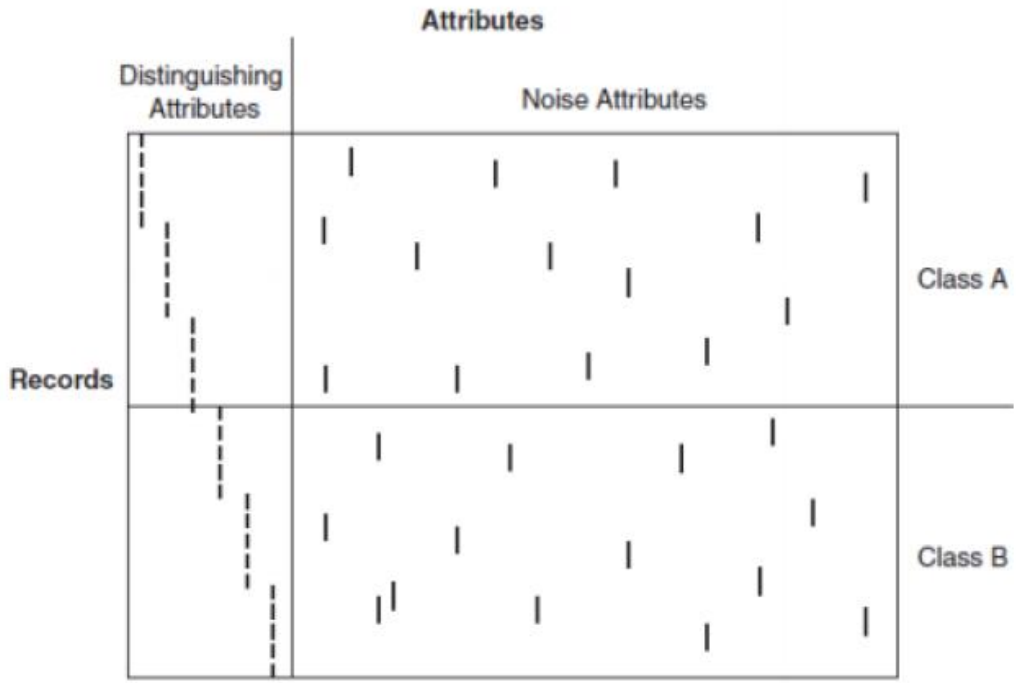
Marital status	Annual income	Default
Single	125K	No
Married	95K	No
Single	70K	No
Married	120K	No
Divorced	75K	Yes
Married	60K	No
Divorced	220K	No
Single	85K	Yes
Married	75K	No
Single	90K	Yes

Assume that the following data set contains 1000 records with class label A and 1000 records with label B. There are some binary variables with distinctive power: $X_1, X_2, \ldots,$ in addition, there are many noisy binary attributes, that take value 1 or 0 at random.

1. Sketch a decision tree that learns such data! How would the decision tree classifier perform on this data?

2. Which records are close to the first record? How would the kNN classifier perform on this data?

3. Outline the conditional probabilities! How would the Naive Bayes classifier perform on this data? Consider the first row as an example!

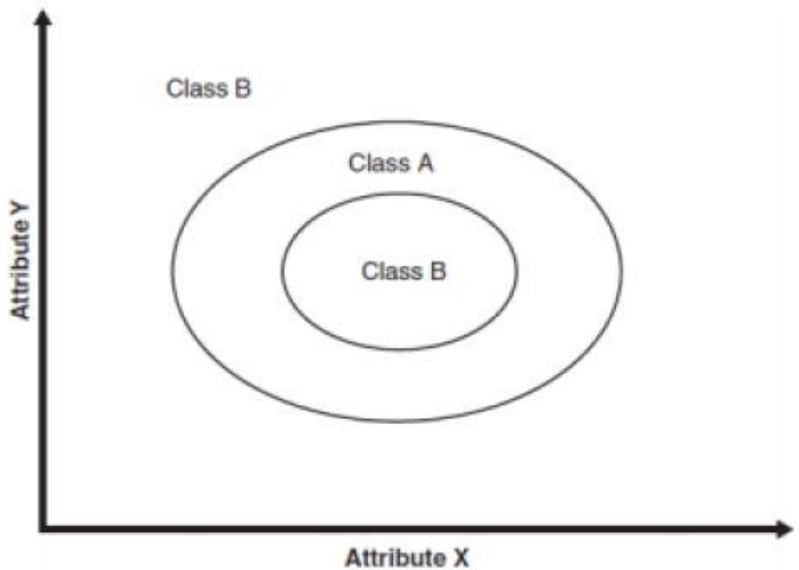


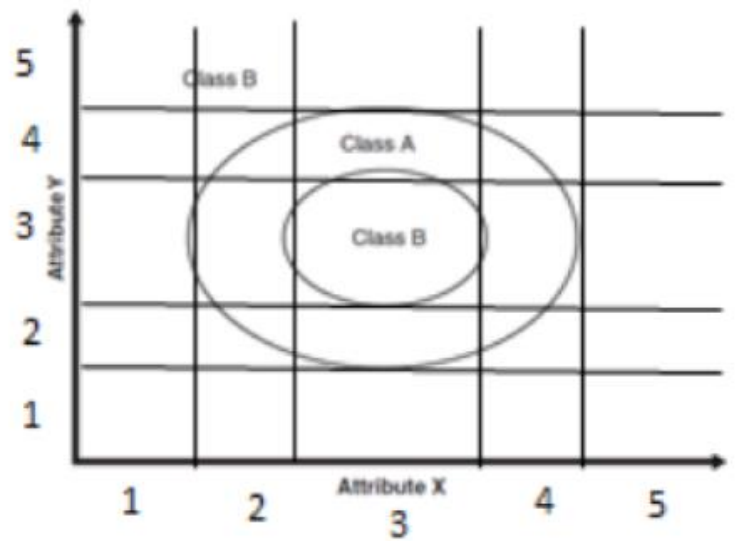
Consider the following data with two attributes (X and Y) and two possible labels (A and B). The position of class A and class B records in the X-Y space is illustrated below.

1. How would a decision tree work on such data? Indicate decision boundaries!

2. How would the kNN classifier perform on this data? What does its performance depend on?

3. How would the Naive Bayes classifier perform? Outline the conditional probabilities! We assume that the two classes have the same number of records and that the instances are distributed uniformly. Use the possible discretization given below, i.e. both attribute X and Y are discretized into 5 bins!





Acknowledgement

- András Benczúr, Róbert Pálovics, SZTAKI-AIT, DM1-2
- Krisztián Buza, MTA-BME, VISZJV68
- Bálint Daróczy, SZTAKI-BME, VISZAMA01
- Judit Csima, BME, VISZM185
- Gábor Horváth, Péter Antal, BME, VIMMD294, VIMIA313
- Lukács András, ELTE, MM1C1AB6E
- Tim Kraska, Brown University, CS195
- Dan Potter, Carsten Binnig, Eli Upfal, Brown University, CS1951A
- Erik Sudderth, Brown University, CS142
- Joe Blitzstein, Hanspeter Pfister, Verena Kaynig-Fittkau, Harvard University, CS109
- Rajan Patel, Stanford University, STAT202
- Andrew Ng, John Duchi, Stanford University, CS229