Review 4: Non-neural Classification Algorithms

한양대학교 ERICA 소프트웨어융합대학 COLLEGE OF COMPUTING | 인공지능학과 Department of Artificial Intelligence

정 우 환 (whjung@hanyang.ac.kr)

Fall 2021

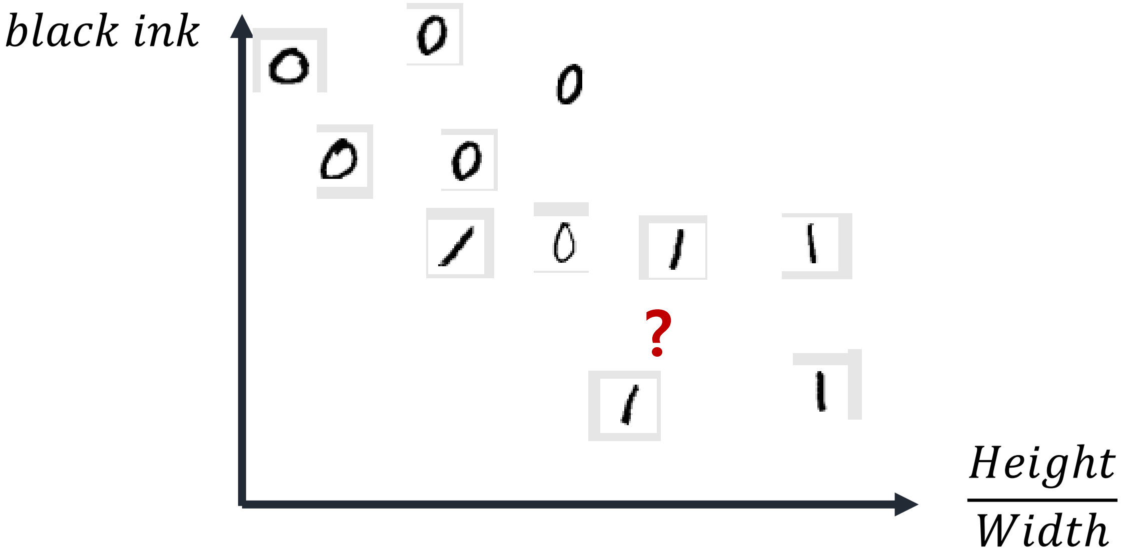
Non-neural classification algorithms

- K-nearest neighbor (k-NN) classifier
- Naïve Bayes classifiers
- Decision trees
- Support Vector Machine (SVM)

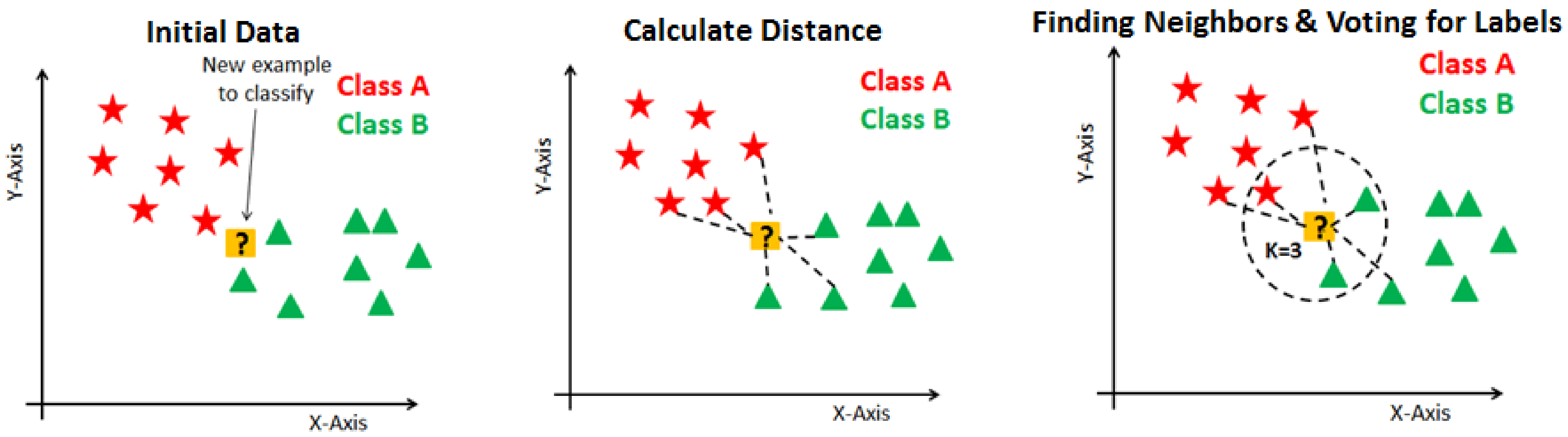
K-NN classifier

K-NN classifier

- K-nearest neighbor classifier (k-NN)
- A non-parametric classification method
- An intuition of KNN



K-NN classifier



https://medium.com/30-days-of-machine-learning/day-3-k-nearest-neighbors-and-bias-variance-tradeoff-75f84d515bdb

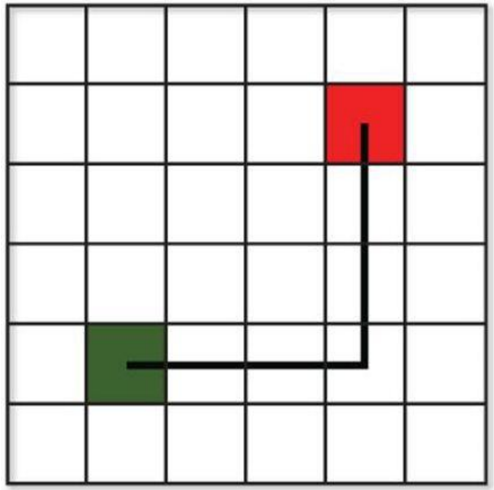
Distance metrics

- A distance function d is **metric** if the following three conditions are satisfied
  - Non-negativity: $d(\boldsymbol{x}, \boldsymbol{y}) \geq 0$
  - Identity: $d(\boldsymbol{x}, \boldsymbol{y}) = 0$ iff $\boldsymbol{x} = \boldsymbol{y}$
  - Symmetry: $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$

- Example) L_p distance $\left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$

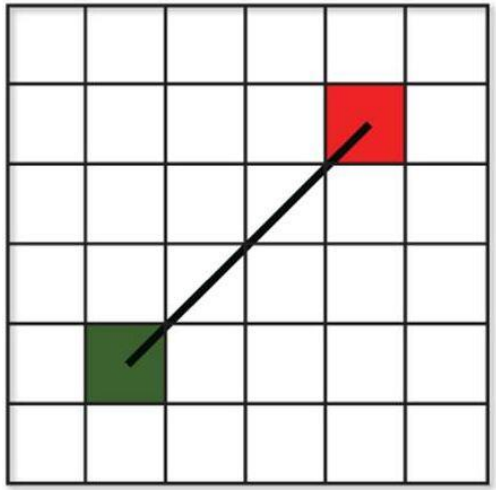
Lp-distances

$$\left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$



Manhattan Distance

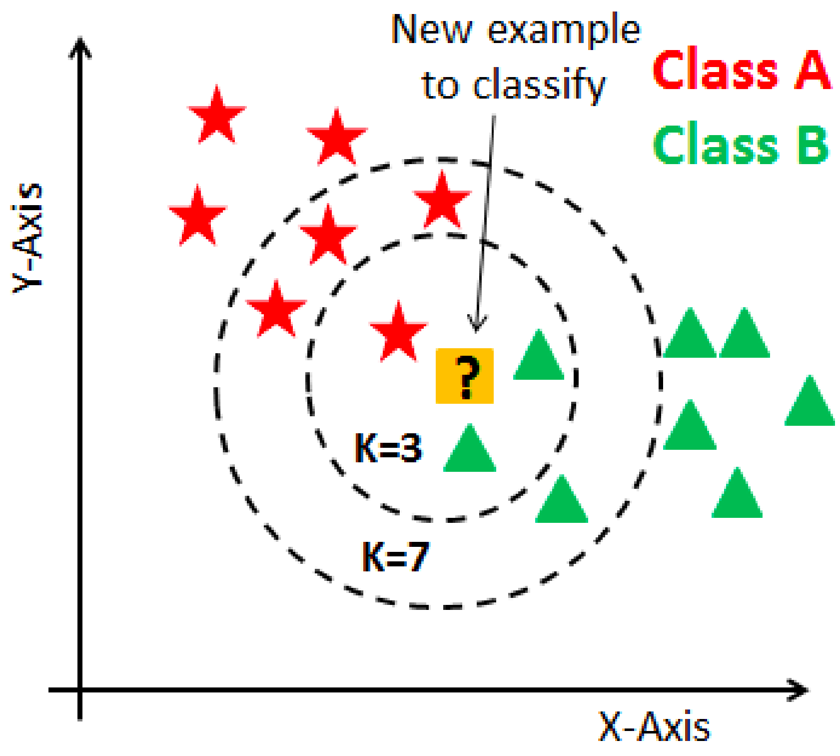
P=1



Euclidean Distance

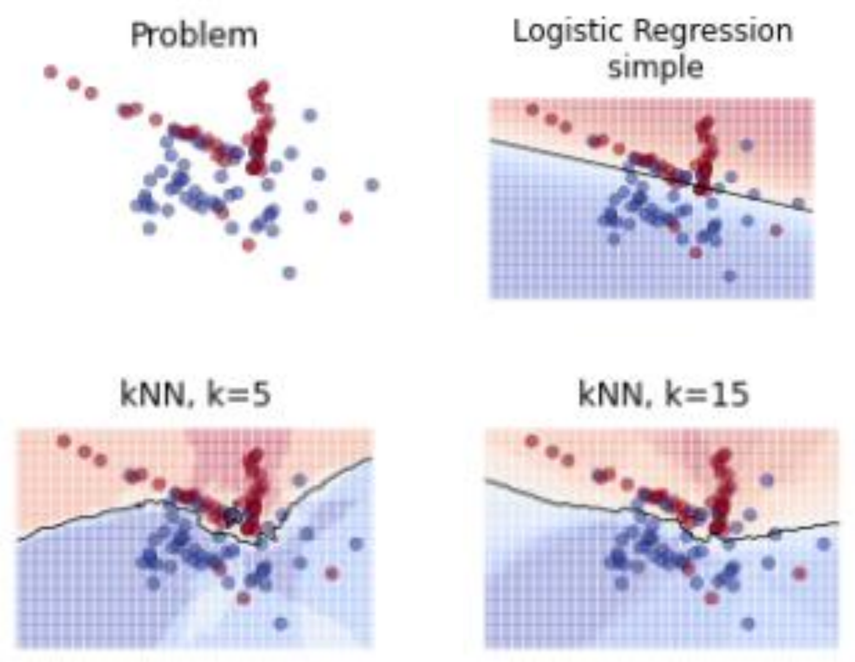
P=2

Different K could have different results



https://medium.com/machine-learning-researcher/k-nearest-neighbors-in-machine-learning-e794014abd2a

Simple, but works well



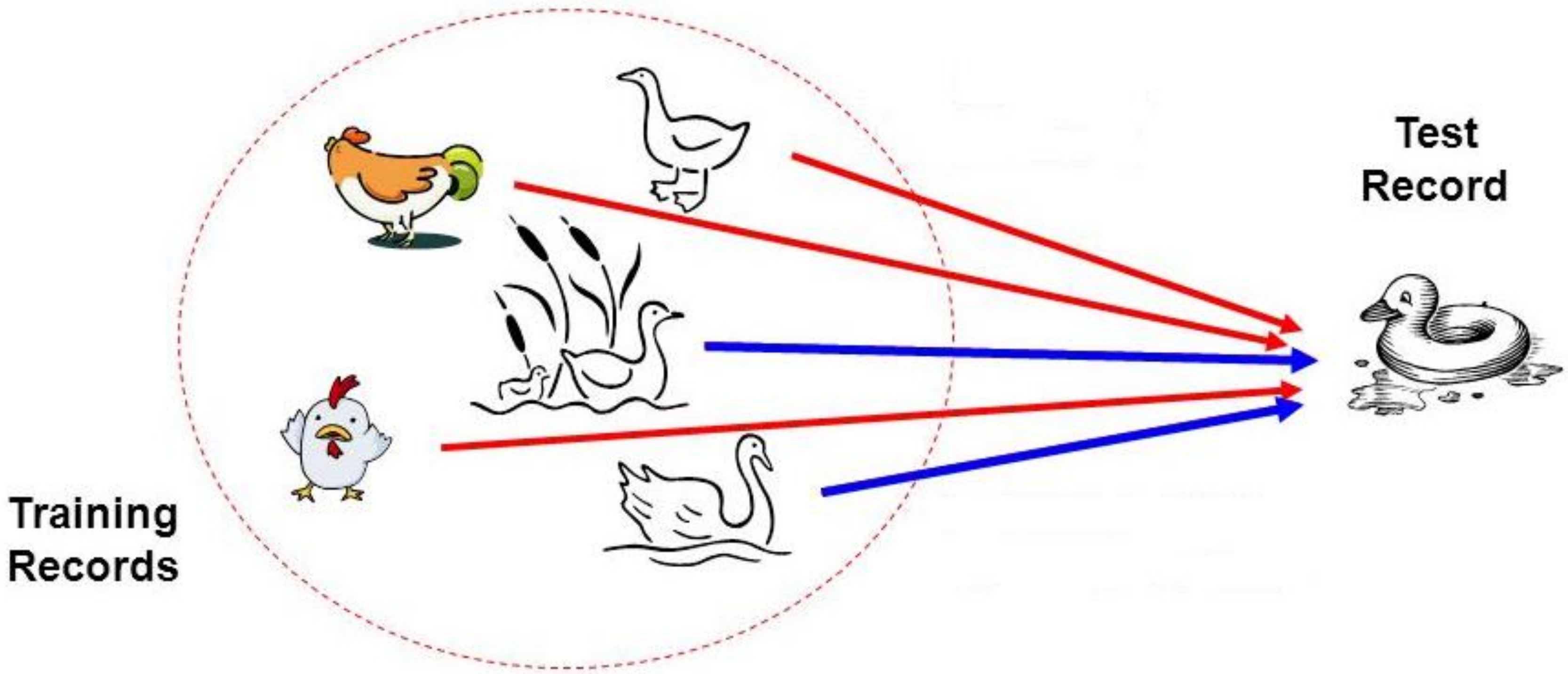
- Some disadvantages of KNN
  - Accuracy depends on the quality of the data
  - Computationally expensive
  - Sensitive to the scale of the data and irrelevant features

Naïve Bayes Classifier

Modified from Prof. Debasis Samanta's slides

Bayesian Classifier

- Principle
  - If it walks like a duck, quacks like a duck, then it is probably a duck



Bayesian Classifier

- A statistical classifier
  - Performs probabilistic prediction, i.e., predicts class membership probabilities
    - Output $p(C_1), p(C_2) \dots p(C_k)$

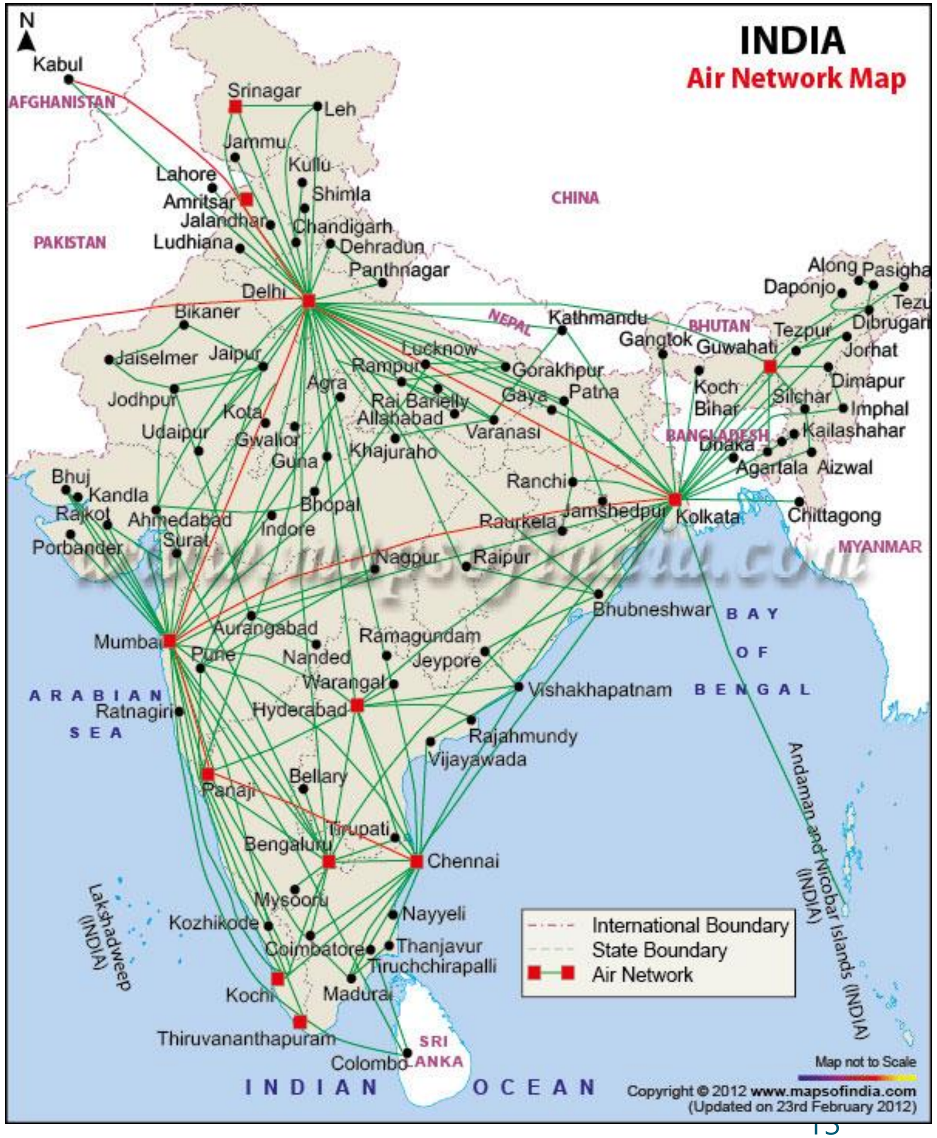
- Foundation
  - Based on Bayes' Theorem.

- Assumptions
  1. The classes are mutually exclusive and exhaustive.
  2. The attributes are independent given the class.

- Called "Naïve" classifier because of these assumptions.
  - Empirically proven to be useful.
  - Scales very well.

Example: Bayesian Classification

- **Example 8.2:** Air Traffic Data
  - Let us consider a set observation recorde d in a database
    - Regarding the arrival of airplanes in the rout es from any airport to New Delhi under cert ain conditions.



Air-Traffic Data

Days	Season	Fog	Rain	Class
Weekday	Spring	None	None	On Time
Weekday	Winter	None	Slight	On Time
Weekday	Winter	None	None	On Time
Holiday	Winter	High	Slight	Late
Saturday	Summer	Normal	None	On Time
Weekday	Autumn	Normal	None	Very Late
Holiday	Summer	High	Slight	On Time
Sunday	Summer	Normal	None	On Time
Weekday	Winter	High	Heavy	Very Late
Weekday	Summer	None	Slight	On Time

Cond. to next slide…

Air-Traffic Data

Cond. from previous slide…

Days	Season	Fog	Rain	Class
Saturday	Spring	High	Heavy	Cancelled
Weekday	Summer	High	Slight	On Time
Weekday	Winter	Normal	None	Late
Weekday	Summer	High	None	On Time
Weekday	Winter	Normal	Heavy	Very Late
Saturday	Autumn	High	Slight	On Time
Weekday	Autumn	None	Heavy	On Time
Holiday	Spring	Normal	Slight	On Time
Weekday	Spring	Normal	None	On Time
Weekday	Spring	Normal	Heavy	On Time

Air-Traffic Data

- In this database, there are four attributes

$$A = [\text{ Day, Season, Fog, Rain}]$$

  with 20 tuples.
- The categories of classes are:

$$C = [\text{On Time, Late, Very Late, Cancelled}]$$

- Given this is the knowledge of data and classes, we are to find most likely classification for any other unseen instance, for example:

Week Day	**Winter**	**High**	**None**	**???**

- Classification technique eventually to map this tuple into an accurate class.

Bayesian Classifier

- In many applications, the relationship between the attributes set and the class variable is non-deterministic.
  - In other words, a test cannot be classified to a class label with certainty.
  - In such a situation, the classification can be achieved probabilistically.

- The Bayesian classifier is an approach for modelling probabilistic relationships between the attribute set and the class variable.

- More precisely, Bayesian classifier use Bayes' Theorem of Probability for classification.

- Before going to discuss the Bayesian classifier, we should have a quick look at the Bayes' Theorem.

Bayes' Theorem

What you know?	What you want to know?
$P(E\|F)$	$P(F\|E)$
$P(Test\ result\|Disease)$	$P(Disease\|Test\ result)$
$P(Power\|Fault)$	$P(Fault\|Power)$
$P(Weather\|Delay)$	$P(Delay\|Weather)$

Bayes Theorem

Want $P(F|E)$, Know $P(E|F)$

- For any events E and F where P(E)>0 and P(F)>0

Likelihood Prior

Posterior $$P(F|E) = \frac{P(E|F)P(F)}{P(E)}$$

- Proof)

$$P(F|E) = \frac{P(EF)}{P(E)}$$ Conditional probability

$$= \frac{P(E|F)P(F)}{P(E)}$$ Chain rule

Naïve Bayesian Classifier

$$\boldsymbol{X} = (x_1, x_2, \dots, x_k) \xrightarrow{\text{Naïve Bayesian Classifier}} y \in \{C_1, C_2, \dots, C_m\}$$

- Classification is to derive the **maximum posteriori**, i.e., the maximal $P(C_i|\boldsymbol{X})$
- This can be derived from **Bayes' theorem**

$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})}$$

- Since P(X) is constant for all classes, only needs to be maximized

$$P(C_i|\mathbf{X}) \propto P(\mathbf{X}|C_i)P(C_i)$$

- Due to the independent assumption,

$$P(C_i|\mathbf{X}) \propto P(\mathbf{X}|C_i)P(C_i) = \left(\prod_{j=1}^{k} P(x_j|C_i)\right)P(C_i)$$

Naïve Bayesian Classifier

- Given $\boldsymbol{X}$
- Output C_i which has the maximum $P(\boldsymbol{X}|C_i)P(C_i) = \left(\prod_{j=1}^{k} P\left(x_j|C_i\right)\right)P(C_i)$
- Example)

Week Day	**Winter**	**High**	**Heavy**	**???**

$$P([WeekDay, Winter, High, Heavy]|On\ time)$$

$$= P(WeekDay|On\ time)P(Winter|On\ time)P(High|On\ time)P(On\ time)$$

Naïve Bayesian Classifier

$$P(\boldsymbol{X}|C_i)P(C_i) = \left(\prod_{j=1}^{k} P(x_j|C_i)\right) P(C_i)$$

- **Example:** With reference to the Air Traffic Dataset mentioned earlier, let us tabulate all the posterior and prior probabilities as shown below.

		Class			
Attribute		On Time	Late	Very Late	Cancelled
Day	Weekday	9/14 = 0.64	½ = 0.5	3/3 = 1	0/1 = 0
	Saturday	2/14 = 0.14	½ = 0.5	0/3 = 0	1/1 = 1
	Sunday	1/14 = 0.07	0/2 = 0	0/3 = 0	0/1 = 0
	Holiday	2/14 = 0.14	0/2 = 0	0/3 = 0	0/1 = 0
Season	Spring	4/14 = 0.29	0/2 = 0	0/3 = 0	0/1 = 0
	Summer	6/14 = 0.43	0/2 = 0	0/3 = 0	0/1 = 0
	Autumn	2/14 = 0.14	0/2 = 0	1/3= 0.33	0/1 = 0
	Winter	2/14 = 0.14	2/2 = 1	2/3 = 0.67	0/1 = 0

Naïve Bayesian Classifier

$$P(\boldsymbol{X}|C_i)P(C_i) = \left(\prod_{j=1}^{k} P(x_j|C_i)\right)P(C_i)$$

		Class			
Attribute		On Time	Late	Very Late	Cancelled
Fog	None	5/14 = 0.36	0/2 = 0	0/3 = 0	0/1 = 0
	High	4/14 = 0.29	1/2 = 0.5	1/3 = 0.33	1/1 = 1
	Normal	5/14 = 0.36	1/2 = 0.5	2/3 = 0.67	0/1 = 0
Rain	None	5/14 = 0.36	1/2 = 0.5	1/3 = 0.33	0/1 = 0
	Slight	8/14 = 0.57	0/2 = 0	0/3 = 0	0/1 = 0
	Heavy	1/14 = 0.07	1/2 = 0.5	2/3 = 0.67	1/1 = 1
Prior Probability		14/20 = 0.70	2/20 = 0.10	3/20 = 0.15	1/20 = 0.05

Naïve Bayesian Classifier

$$P(\boldsymbol{X}|C_i)P(C_i) = \left(\prod_{j=1}^{k} P(x_j|C_i)\right) P(C_i)$$

Instance:

Week Day	Winter	High	Heavy	???

Case1: Class = On Time : 0.70 × 0.64 × 0.14 × 0.29 × 0.07 = 0.0013

Case2: Class = Late : 0.10 × 0.50 × 1.0 × 0.50 × 0.50 = 0.0125

Case3: Class = Very Late : 0.15 × 1.0 × 0.67 × 0.33 × 0.67 = 0.0222

Case4: Class = Cancelled : 0.05 × 0.0 × 0.0 × 1.0 × 1.0 = 0.0000

Case3 is the strongest; Hence correct classification is **Very Late**

Naïve Bayesian Classifier

Pros and Cons

- The Naïve Bayes' approach is a very popular one, which often works well.
- However, it has a number of potential problems
  - It relies on all attributes being categorical.
  - If the data is less, then it estimates poorly.
  - …

Naïve Bayesian Classifier

Approach to overcome the limitations in Naïve Bayesian Classification

- Estimating the posterior probabilities for continuous attributes
  - In real life situation, all attributes are not necessarily be categorical, In fact, there is a mix of both categorical and continuous attributes.
  - In the following, we discuss the schemes to deal with continuous attributes in Bayesian classifier.

1. Discretize each continuous attribute and then replace the continuous values with its corresponding discrete intervals.

$$24.3°C \rightarrow [20°C, 25°C)$$

2. Assume a certain form of probability distribution for the continuous variable
   Gaussian distribution is widely used

$$P(x: \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where, μ and σ^2 denote mean and variance, respectively.

Naïve Bayesian Classifier

$$P(\boldsymbol{X}|C_i)P(C_i) = \left(\prod_{j=1}^{k} P\left(x_j|C_i\right)\right)P(C_i)$$

- For each class C_i and attribute j,
  - $P(x_j|\mathrm{C}_i) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{-\frac{\left(x_j-\mu_{ij}\right)^2}{2\sigma_{ij}{}^2}}$
    - μ_{ij}: Sample mean
    - σ_{ij}^2: Sample variance

Additive Smoothing

Naïve Bayes

$$P(\boldsymbol{X}|C_i)P(C_i) = \left(\prod_{j=1}^{k} P(x_j|C_i)\right)P(C_i)$$

Approach to overcome the limitations in Naïve Bayesian Classification

$$P(x_j|C_i) = \frac{n_{c_i}}{n}$$

n = total number of instances from class C_i

n_{c_i} = number of training examples from class C_i that take the value $x_j = a_j$

- If the training data size is too small..
  - $P(x_j|C_i) \to 0$ for some ij
- Additive smoothing

$$P(x_j|C_i) = \frac{n_{c_i} + \alpha}{n + m_j\alpha}$$

$m_j\text{: \# of possible values of attribute } i$