University of South Australia

kNN and Naïve Bayes classifiers

Dr Srećko Joksimović

slido

What was the first rule in your practical from Week 7?

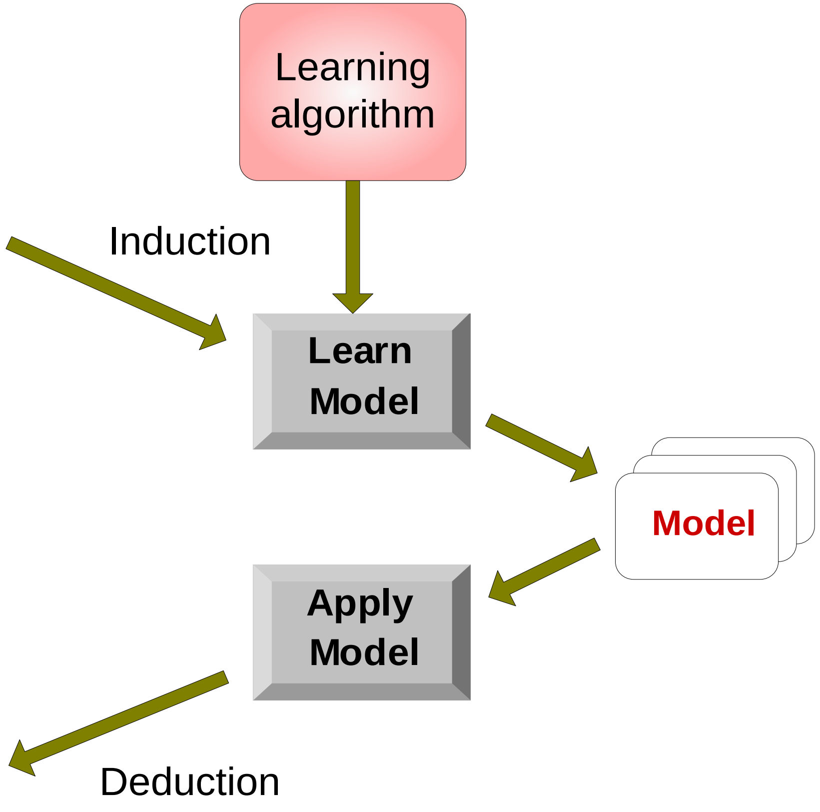
ⓘ Start presenting to display the poll results on this slide.

K-Nearest Neighbour Classifiers

Classification

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set



Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

- Eager learners
  - decision tree, rule-based
- Lazy learners
  - Rote classifier, K Nearest-Neighbor classifier

Image credit: https://ourcodingclub.github.io/tutorials/machine-learning/

KNN

The KNN or k-nearest neighbours algorithm is one of the simplest, distance-based, machine learning algorithms.

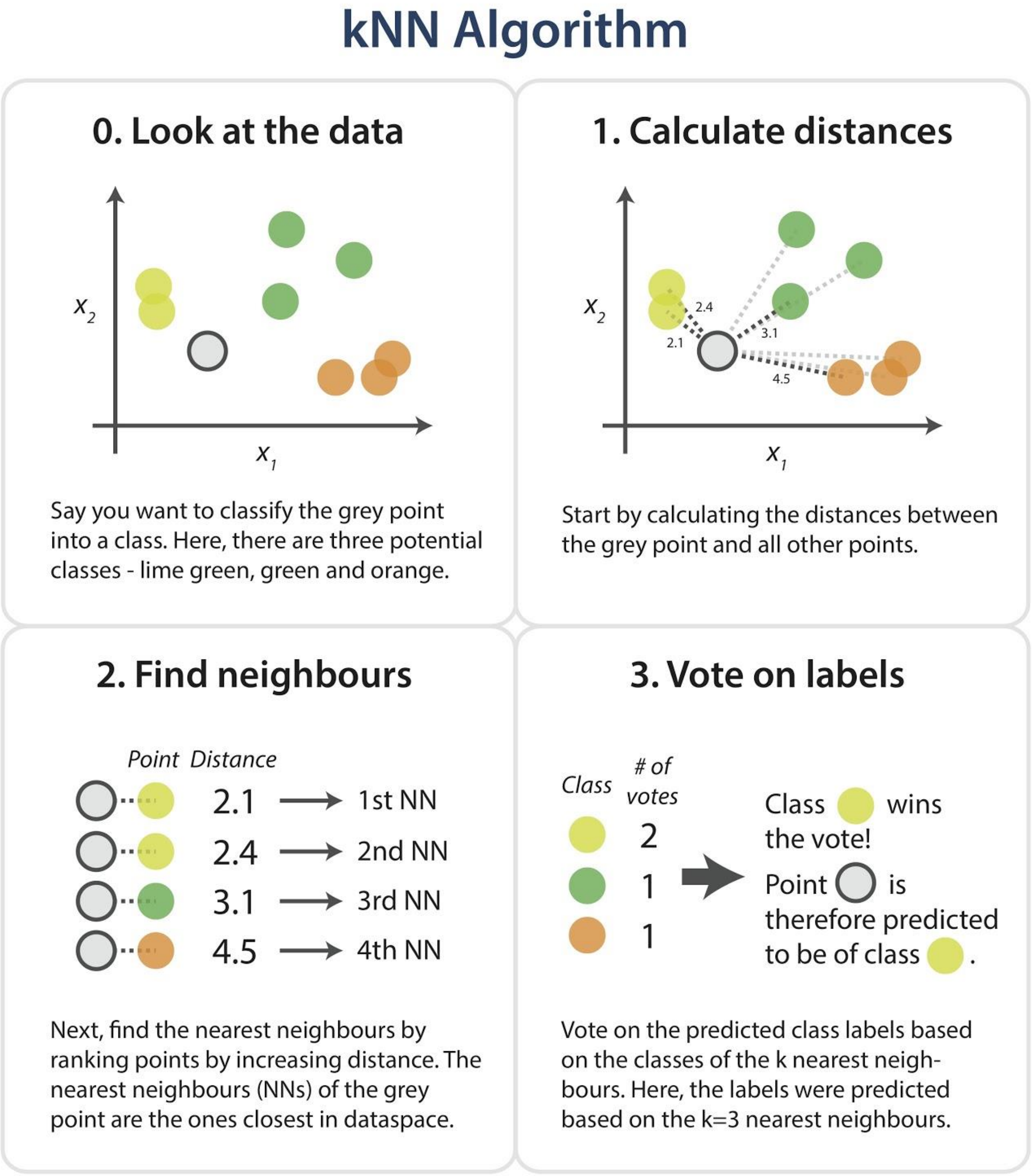
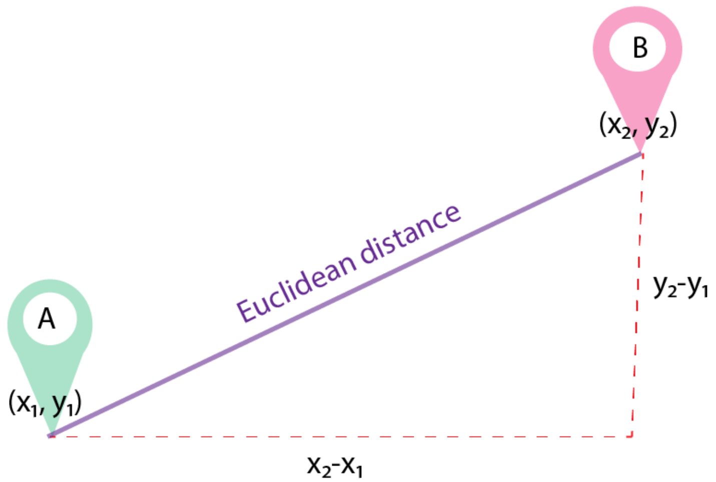


Image credit: https://ourcodingclub.github.io/tutorials/machine-learning/

Euclidian distance

Euclidean distance can simply be defined as the shortest between the 2 points irrespective of the dimensions.

$$\text{dist}(A, B) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$



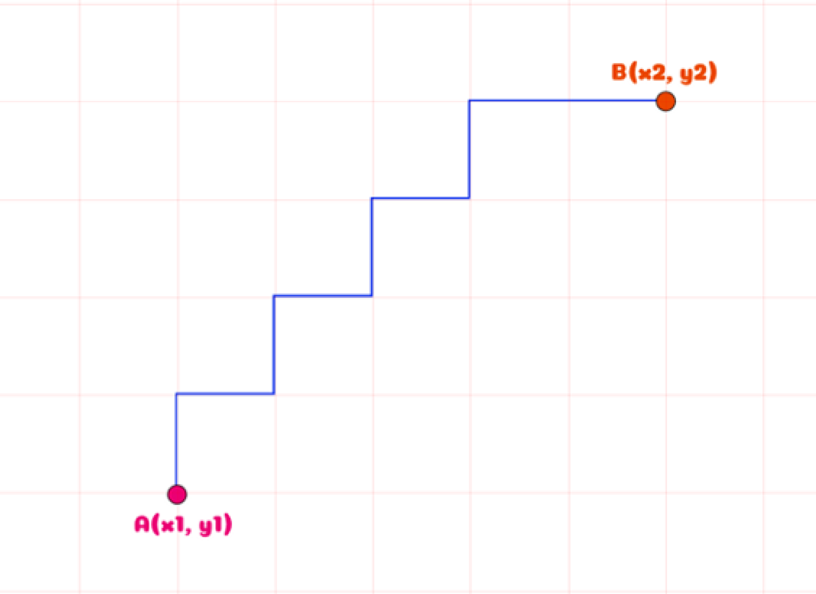
Source: “Different Types of Distance Metrics used in Machine Learning”

Manhattan distance

The distance between two points measured along axes at right angles.

$$\text{dist}(\text{A}, \text{B}) = \sum_{i=1}^{N} |fai \ - fbi|$$

For example, if A=(x1, x2) and B=(y1, y2),
the Manhattan distance between |x1 - y1| + |x2 – y2|



Minkowski distance

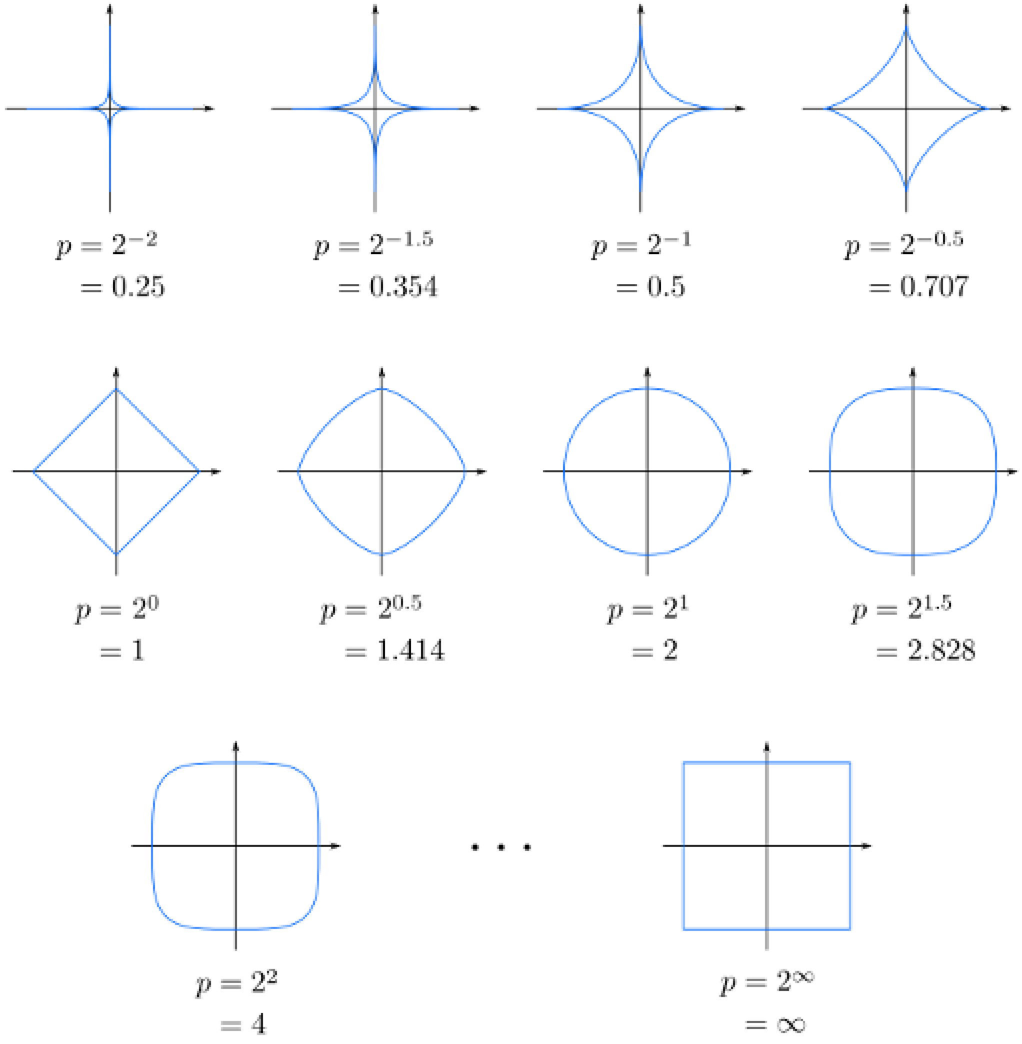
Minkowski distance is a distance measurement between two points in the normed vector space (N-dimensional real space) and is a generalization of the Euclidean distance.

For:

- p = 1, Minkowski = Manhattan
- P = 2, Minkowski = Eucledian

$$\sqrt[p]{(x1 - y1)^p + (x2 - y2)^p + \ldots + (xN - yN)^p}$$

Source: "Different Types of Distance Metrics used in Machine Learning"



Unit circles with various values of p (Minkowski distance)
OpenGenus

Cosine similarity

- Cosine similarity is a measure of similarity between two non-zero vectors.

- Cosine similarity measures the cosine of the angle between the two vectors.

- It is a value between [-1, 1]
  - 1 means the vectors are identical,
  - 0 means they are orthogonal (i.e., unrelated), and
  - -1 means they are diametrically opposed.

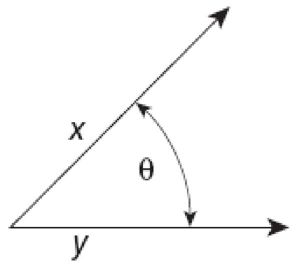


Figure 2.16. Geometric illustration of the cosine measure.

Cosine similarity

- Cosine similarity is a calculated as follows:
  - Given two vectors A and B with n dimensions, their cosine similarity is defined as the dot product of A and B divided by the product of their magnitudes

$$\cos(A, B) = \frac{A \cdot B}{||A||x||B||}$$

where "." denotes the dot product of A and B, and "||A||" (i.e., Euclidian norm) denotes the magnitude of a vector.

$$||A|| = \sqrt{a1^2 + a2^2 + \ \ldots + \ an^2}$$

Cosine similarity - Example

A = 3 2 0 5 0 0 0 2 0 0
B = 1 0 0 0 0 0 0 1 0 2

$$A \cdot B = 3*1+2*0+0*0+5*0+0*0+0*0+0*0+2*1+0*0+0*2 = 5$$

$$||A|| = sqrt(3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0) = 6.48$$

$$||B|| = sqrt(1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2) = 2.24$$

$$\cos(A,B) = \frac{5}{6.48*2.24} = 0.34$$

The choice of distance measure

- **Euclidean distance**: This distance measure is widely used in KNN, especially for problems with **continuous numerical features**.
  - It works well when the data is well-scaled and the features have similar units of measurement.
- **Manhattan distance**: This distance measure is suitable for problems where the data has **categorical features**.
- **Minkowski distance**: It's a generalization of both the Euclidean and Manhattan distances, and is useful when **the degree of the distance measure needs to be tuned** to better match the underlying data distribution.
- **Cosine similarity:** This distance measure is suitable for problems involving **text or other forms of high-dimensional data**.

Why scaling?

All distance based algorithms are affected by the scale of the variables.

ID	Age	Income ($)
1	25	80,000
2	30	100,000
3	40	90,000
4	30	50,000
5	40	110,000

Eucledian distance (P1, P2) ≈ 20,000

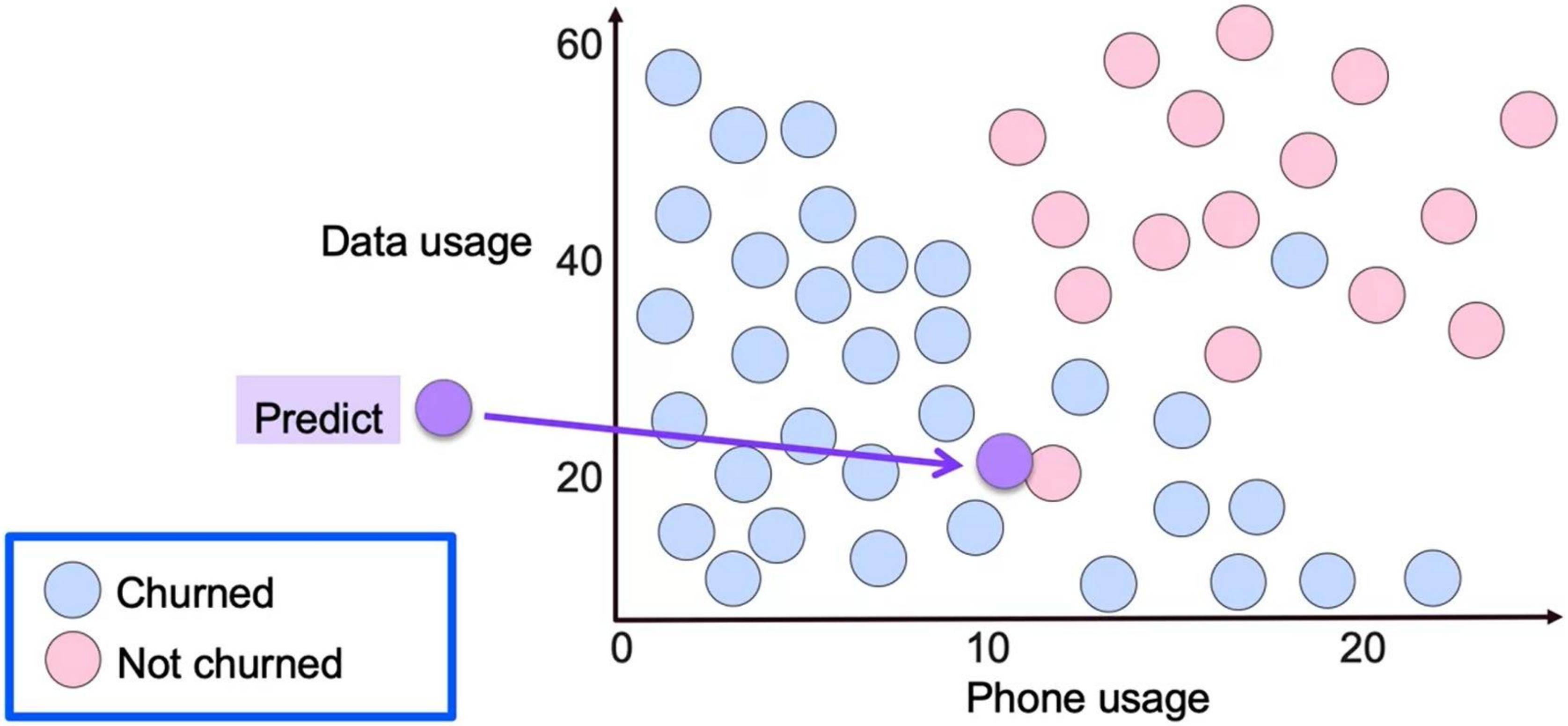
If we normalise the data above, we will get

$$z = \frac{x - \mu}{\sigma}$$

ID	Age	Income ($)
1	-1.192	-0.260
2	-0.447	0.608
3	1.043	0.173
4	-0.447	-1.563
5	1.043	1.042

Eucledian distance (P1, P2) = 1.1438

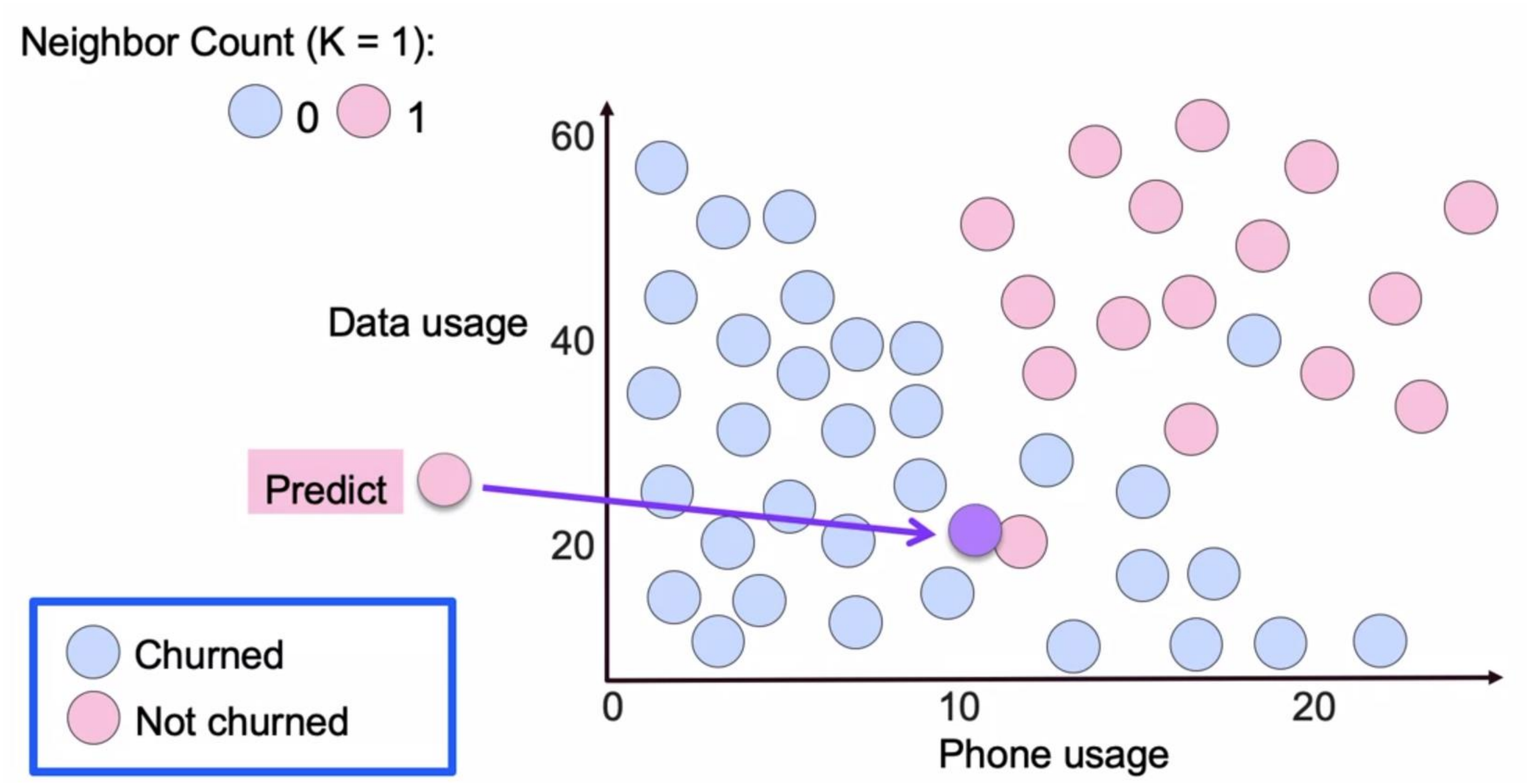
How to choose K – why do we care?



University of
South Australia

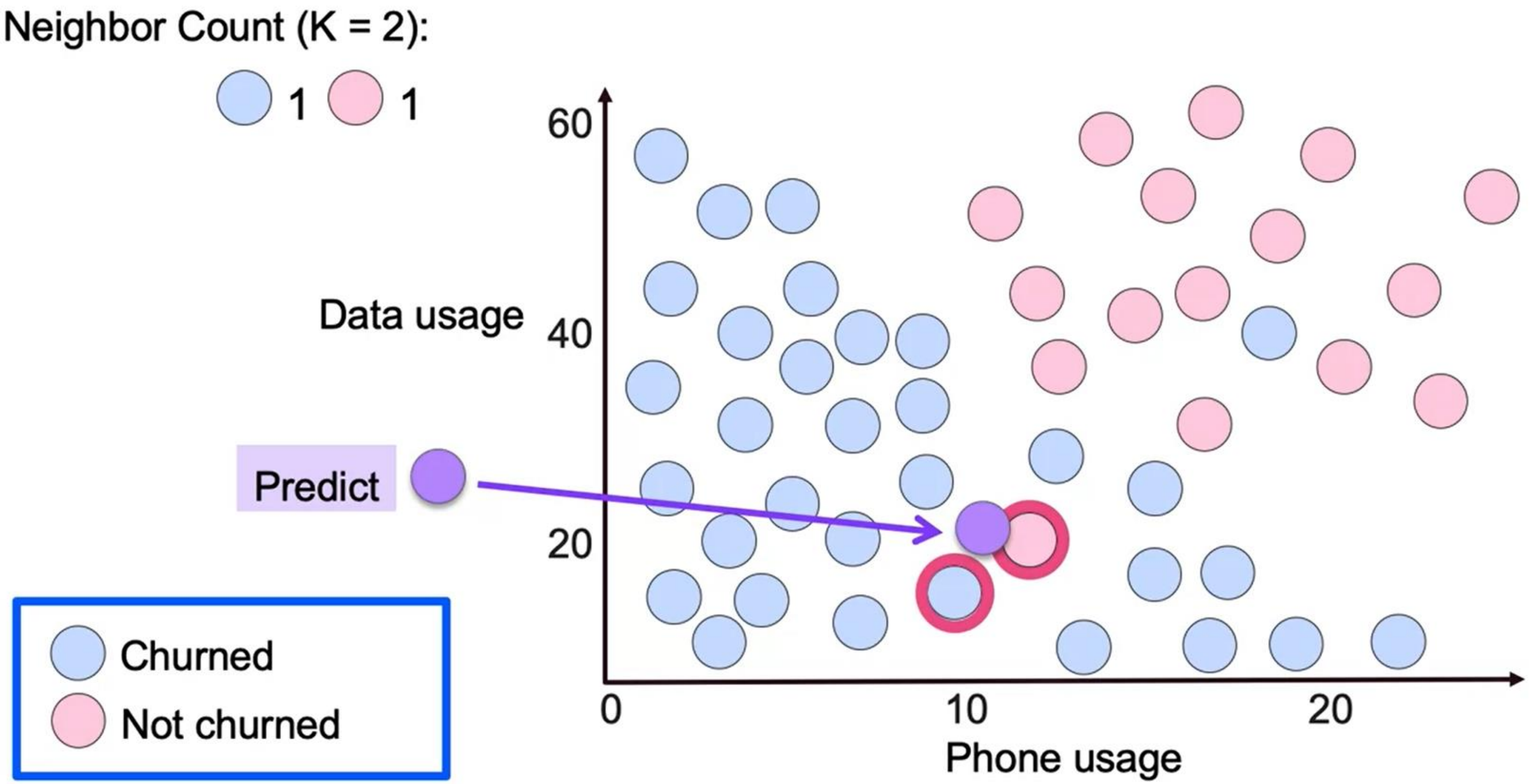
Credit: https://www.coursera.org/learn/supervised-machine-learning-classification

How to choose K – why do we care?



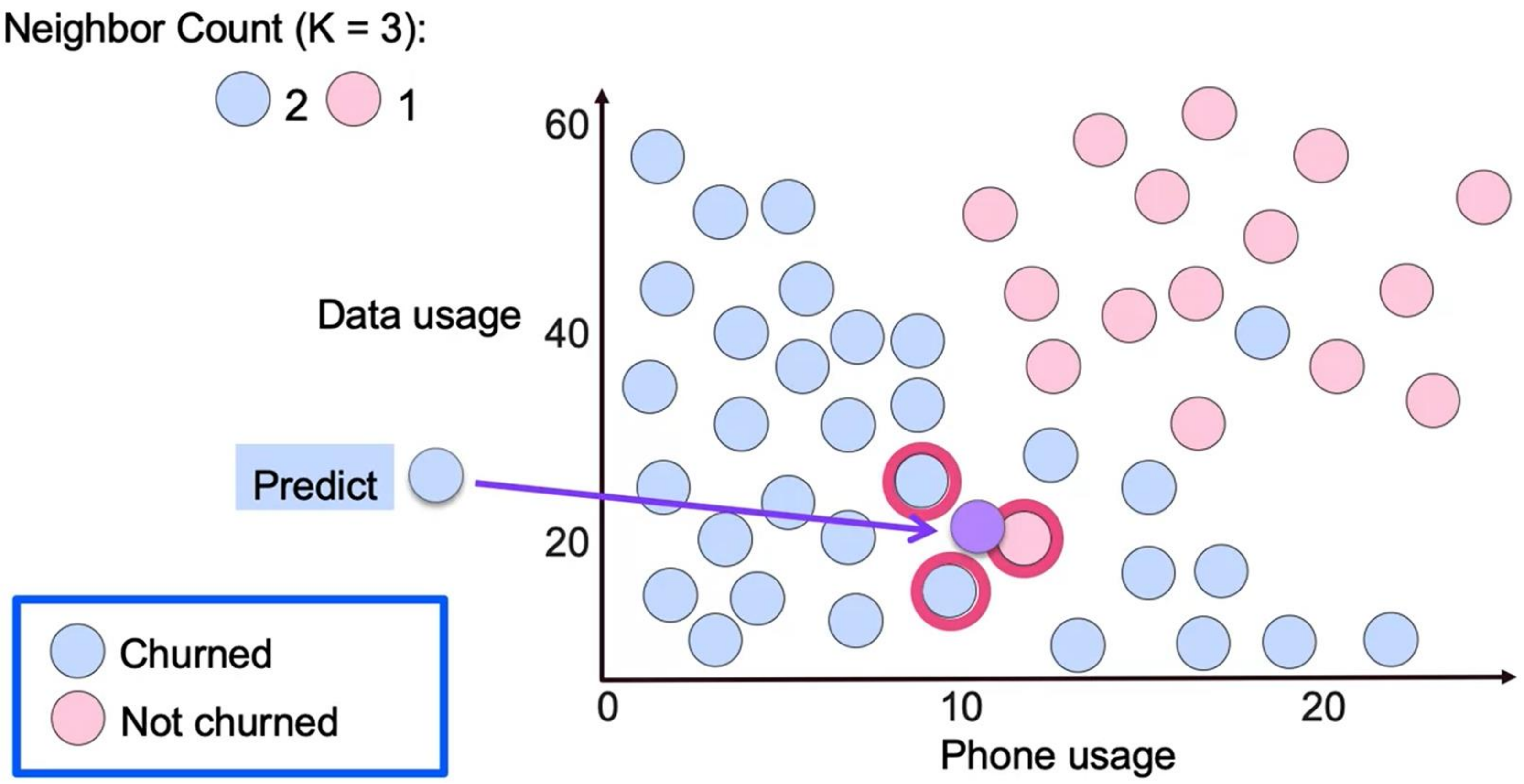
Credit: https://www.coursera.org/learn/supervised-machine-learning-classification

How to choose K – why do we care?



Credit: https://www.coursera.org/learn/supervised-machine-learning-classification

How to choose K – why do we care?



Credit: https://www.coursera.org/learn/supervised-machine-learning-classification

So, how to choose *K*?

There is no “correct” *K*

The right value depends on the metric that is most important

Common approaches:

- Square root of the number of samples
- Elbow plot
- Overfitting estimate (see Week 8 Practical)

Credit: https://www.coursera.org/learn/supervised-machine-learning-classification

Characteristics of Nearest-Neighbor Classifiers

- Instance-based learning – no need to maintain an abstraction (or model)
- No need for model building
- Makes a prediction based on a local information – **susceptible to noise**
- Can produce arbitrarily shaped decision boundaries
- Can produce **wrong predictions** unless appropriate proximity measure and data pre-processing steps are taken

Credit: https://www.coursera.org/learn/supervised-machine-learning-classification

Naïve Bayes Classifiers

Uncertainty and probability

- The Naïve Bayes machine learning algorithm is one of the tools to deal with uncertainty with the help of probabilistic methods.
- Probability is a field of math that enables us to reason about uncertainty and assess the likelihood of some results or events.
- In machine learning, we are interested in **conditional probabilities**. We are interested not in the general probability that something will happen, but the **likelihood it will happen given that something else happens**.
- This is how conditional probability is defined: the probability of Y, given X = the joint probability of both Y and X happening, divided by the probability of X.

$$P(Y|X) = \frac{P(Y \wedge X)}{P(X)}$$

Bayes Theorem

$$P(Y \mid X) = \frac{P(X \mid Y)\, P(Y)}{P(X)}$$

Y, X - events

P(Y | X) – probability of Y given X is true

P(X | Y) – probability of X given Y is true

P(Y), P(X) – probabilities of Y and X

Bayes theorem allows us to calculate conditional probabilities.

It comes extremely handy because it enables us to use some knowledge that we already have (called prior) to calculate the probability of a related event.

It is used in developing models for classification and predictive modeling problems such as Naive Bayes.

Bayes Theorem - Example

Consider a football game between two rival teams: Team A and Team B.
Suppose Team A wins 65% of the time and Team B wins the remaining matches.
Among the games won by Team A, only 30% of them come from playing on Team B's football field. On the other hand, 75% of the victories for Team B are obtained while playing at home. If Team B is to host the next match between two teams, which team will most likely emerge as a winner?

Bayes Theorem - Example

Probability Team A wins is $P(Y=0) = 0.65$

Probability Team B wins is $P(Y=1) = 1 - P(Y=0) = 0.35$

Probability Team B hosted the match it won is $P(X=1|Y=1) = 0.75$

Probability Team B hosted the match won by Team A is $P(X=1|Y=0) = 0.3$

$P(Y=1|X=1)$?

Bayes Theorem - Example

Probability Team A wins is $P(Y=0) = 0.65$
Probability Team B wins is $P(Y=1) = 1 - P(Y=0) = 0.35$
Probability Team B hosted the match it won is $P(X=1|Y=1) = 0.75$
Probability Team B hosted the match won by Team A is $P(X=1|Y=0) = 0.3$

$$P(Y = 1|X = 1) = \frac{P(X = 1|Y = 1)\ x\ P(Y = 1)}{P(X = 1)}$$

Bayes Theorem - Example

Probability Team A wins is $P(Y=0) = 0.65$
Probability Team B wins is $P(Y=1) = 1 - P(Y=0) = 0.35$
Probability Team B hosted the match it won is $P(X=1|Y=1) = 0.75$
Probability Team B hosted the match won by Team A is $P(X=1|Y=0) = 0.3$

$$P(Y = 1|X = 1) = \frac{P(X = 1|Y = 1)\ x\ P(Y = 1)}{P(X = 1)}$$

The law of total probability -> $P(X = 1) = P(X = 1, Y = 0) + P(X = 1, Y = 1)$

The product rule of probability -> $P(X = 1) = P(X = 1|Y = 0)P(Y = 0) + P(X = 1|Y = 1)P(Y = 1)$

Bayes Theorem - Example

Probability Team A wins is $P(Y=0) = 0.65$
Probability Team B wins is $P(Y=1) = 1 - P(Y=0) = 0.35$
Probability Team B hosted the match it won is $P(X=1|Y=1) = 0.75$
Probability Team B hosted the match won by Team A is $P(X=1|Y=0) = 0.3$

The product rule of probability ->

$$P(X = 1) = P(X = 1|Y = 0)P(Y = 0) + P(X = 1|Y = 1)P(Y = 1)$$

$$P(Y = 1|X = 1) = \frac{0.75\ x\ 0.35}{0.3\ x\ 0.65 + 0.75\ x\ 0.35}$$

$$P(Y = 1|X = 1) = 0.58$$

$$P(Y \mid X) = \frac{P(X \mid Y)\, P(Y)}{P(X)}$$

P(X) – constant

P(Y) – a fraction of training records that belong to each class

P(X|Y) – Naïve Bayes classifier or Bayesian Belief network (not covered in this course)

Naïve Bayes classifier

- Naive Bayes is a simple supervised machine learning algorithm that uses the Bayes' theorem with strong independence assumptions between the features to procure results.
- That means that the algorithm just assumes that each input variable is independent. It really is a naive assumption to make about real-world data.
    - "I like Harry Potter", "Harry Potter like I", "Potter I like Harry"
- The algorithm is able to effectively solve many complex problems.

Naïve Bayes classifier

$$P(X|Y = y) = \prod_{i=1}^{d} P(X_i|Y = y)$$

$$X = \{X1, X2, \dots, Xd\}$$

Conditional independence

Let X_1, X_2, and Y denote three sets of random variables.

The variables in X_1 are said to be conditionally independent of X_2, given Y, if the following condition holds:

$$P(X_1 | X_2, Y) = P(X_1 | Y)$$

Conditional independence

$$P(X1, X2|Y) = \frac{P(X1, X2, Y)}{P(Y)}$$

$$= \frac{P(X1, X2, Y)}{P(X2, Y)} \text{ x } \frac{P(X2, Y)}{P(Y)}$$

$$= P(X1|X2, Y) \text{ x } P(X2, Y)$$

$$= P(X1|Y) \text{ x } P(X2|Y)$$

How Naïve Bayes works?

- Estimate the conditional probability of each X_i, given Y

MAX

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{d} P(X_i|Y)}{P(X)}$$

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, P(Y=y)\prod_{i=1}^{d} P(X_i|Y=y)$$

How Naïve Bayes works? (Example)

Step 1: Convert the dataset into a frequency table

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Source: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained

How Naïve Bayes works? (Example)

Step 2: Create likelihood table by finding the probabilities

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

Likelihood Table 2

Whether	No	Yes	Posterior Probability for No	Posterior Probability for Yes
Overcast		4	0/5=0	4/9=0.44
Sunny	2	3	2/5=0.4	3/9=0.33
Rainy	3	2	3/5=0.6	2/9=0.22
Total	5	9		

Source: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained

How Naïve Bayes works? (Example)

Step 3: Use Naïve Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.

Source: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained

How Naïve Bayes works? (Example)

Problem: Players will play if weather is sunny.

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

Likelihood Table 2

Whether	No	Yes	Posterior Probability for No	Posterior Probability for Yes
Overcast		4	0/5=0	4/9=0.44
Sunny	2	3	2/5=0.4	3/9=0.33
Rainy	3	2	3/5=0.6	2/9=0.22
Total	5	9		

$$P(Yes\ |Sunny) = \frac{P(Sunny\,|\,Yes)\ P(Yes)}{P(Sunny)}$$

P(Sunny |Yes) = 3/9 = 0.33,
P(Sunny) = 5/14 = 0.36,
P(Yes)= 9/14 = 0.64

P (Yes | Sunny) = 0.33 * 0.64 / 0.36 = 0.60

Source: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained

Naïve Bayes advantages

- Easy implementation
- Fast and simple
- Scaling advantages
- Noise resilience
- Easy training
- No overfitting
- Computationally efficient
- Suitable for large dataset

Naïve Bayes disadvantages

- Real world problems
- No regression
- Limited application case
- Biased nature

University of South Australia

INFS 5100 Predictive Analytics

Q&A