EE 4146 Data Engineering and Learning Systems

Lecture 11: Bayes Classifier and KNN

Semester A, 2021-2022

Schedules

Week	Date	Topics
1	Sep. 1	Introduction
2	Sep. 8	Data exploration
3	Sep. 15	Feature reduction and selection (HW1 out)
4	Sep. 22	Mid-Autumn Festival
5	Sep. 29	Clustering I: Kmeans based models (HW1 due in this weekend)
6	Oct. 6	Clustering II: Hierarchical/density based/fuzzing clustering
7	Oct. 13	Midterm (no tutorials this week)
8	Oct. 20	Adverse Weather
9	Oct. 27	Linear classifiers
10	Nov. 3	Classification based on decision tree (Tutorial on project) (HW2 out)
11	Nov. 10	Bayes based classifier and KNN (Tutorial on codes) (HW2 due in this weekend)
12	Nov. 17	Classifier ensemble
13	Nov. 24	Deep learning based models (Quiz)
14	Make up To be decided later	Summary

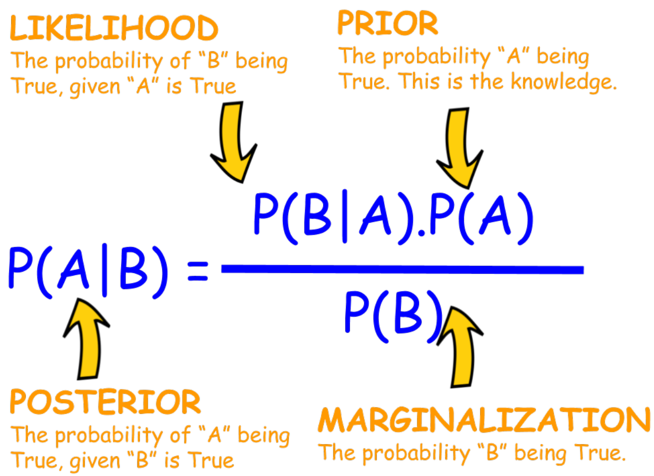
Quiz 2

- Zoom Quiz
  - We will have the second quiz on Nov. 24 from 4:00 PM-5:00 PM. You will have 15 mins to scan your results and uploaded them through the assignment. It will have 4 calculation & understanding-related questions, covering all lecture notes (more emphasis on the notes after the midterm). Please join in this quiz through Canvas ZOOM.
  - Please take photos of your hand-written results together with your Cityu id, combine these calculation results in one file, and upload the file through the assignment. As mentioned in the class, you can use the matlab to calculate the final results, but you should write the detailed steps for each question. The one with only final results will not get the full marks.
  - No code-related questions.
  - Open-book and open-notes.
  - No late submission is allowed since you have 15mins to upload your results. If you could not upload the results, please send me emails of your results.
- Lecture at 5:30 PM- 6:50 PM, Nov. 24

- Bayes based classifier
- KNN
- Review of Decision Tree

Bayes Classifier

- A probabilistic framework for solving classification problems



- Conditional Probability:

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

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

- Bayes theorem:

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

Example of Bayes Theorem

- Given:
  - A doctor knows that meningitis (disease1) causes stiff neck 50% of the time
  - Prior probability of any patient having meningitis is 1/50,000
  - Prior probability of any patient having stiff neck is 1/20

- If a patient has stiff neck, what's the probability he/she has meningitis?
  - Let us formulate this question as a math formula.
  - Let the event of having a stiff neck be S. Let the event of having meningitis be M, what formula can describe this question?

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

Using Bayes Theorem for Classification

- Consider each attribute and class label as random variables
- Given a record with attributes (X_1, X_2,…, X_d)
  - Goal is to predict class Y
  - Specifically, we want to find the value of Y that maximizes P(Y| X_1, X_2,…, X_d)
- Can we estimate P(Y| X_1, X_2,…, X_d) directly from data?

Example Data

- Given a Test Record:

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

$$X = (\mathrm{Refund} = \mathrm{No}, \mathrm{Divorced}, \mathrm{Income} = 120\mathrm{K})$$

- Can we estimate

  P(Evade = Yes | X) and P(Evade = No | X)?

  In the following we will replace

  “Evade = Yes” by Yes, and

  “Evade = No” by No

Using Bayes Theorem for Classification

- Approach:
  - compute posterior probability $P(Y \mid X_1, X_2, \ldots, X_d)$ using the Bayes theorem

$$\boxed{P(Y \mid X_1 X_2 \ldots X_n) = \frac{P(X_1 X_2 \ldots X_d \mid Y) P(Y)}{P(X_1 X_2 \ldots X_d)}}$$

  - *Maximum a-posteriori*: Choose Y that maximizes
    $P(Y \mid X_1, X_2, \ldots, X_d)$ (e.g. Y=yes or Y=no)

  - Equivalent to choosing value of Y that maximizes
    $P(X_1, X_2, \ldots, X_d \mid Y)\, P(Y)$

- How to estimate $P(X_1, X_2, \ldots, X_d \mid Y)$?

Using Bayes Theorem for Classification

$$\boxed{P(Y \mid X_1 X_2 \ldots X_n) = \frac{P(X_1 X_2 \ldots X_d \mid Y) P(Y)}{P(X_1 X_2 \ldots X_d)}}$$

- *Maximum a-posteriori*: Choose Y that maximizes
  $P(Y \mid X_1, X_2, \ldots, X_d)$

- E.g. P(Y=iris Versicolor| petal length=2, petal width=1.5, sepal width=1.4, sepal length=3) =?
- P(Y=iris Setosa| petal length=2, petal width=1.5, sepal width=1.4, sepal length=3) =?
- P(Y=iris Virginica| petal length=2, petal width=1.5, sepal width=1.4, sepal length=3) =?

Example Data

- Given a Test Record:

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

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

Using Bayes Theorem:

- $P(\text{Yes} \mid X) = \dfrac{P(X \mid \text{Yes})P(\text{Yes})}{P(X)}$
- $P(\text{No} \mid X) = \dfrac{P(X \mid \text{No})P(\text{No})}{P(X)}$
- How to estimate $P(X \mid \text{Yes})$ and $P(X \mid \text{No})$?

Naïve Bayes on Example Data

- Given a Test Record:

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

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

- P(X | Yes) =

  P(Refund = No | Yes) x

  P(Divorced | Yes) x

  P(Income = 120K | Yes)

- P(X | No) =

  P(Refund = No | No) x

  P(Divorced | No) x

  P(Income = 120K | No)

Estimate Probabilities from Data

- Class: $P(Y) = N_c/N$
  - e.g., $P(No) = 7/10$,
    $P(Yes) = 3/10$

- For categorical attributes:

  $$P(X_i \mid Y_k) = |X_{ik}| / N_{ck}$$

  - where $|X_{ik}|$ is number of instances having attribute value X_i and belonging to class Y_k
  - Examples:

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

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**

Estimate Probabilities from Data

- For **continuous attributes**:
  - Discretization: Partition the range into bins
  - Replace continuous value with bin value
  - Attribute changed from continuous to ordinal
- Probability density estimation:
  - Assume attribute follows a normal distribution
  - Use data to estimate parameters of distribution
    (e.g., mean and standard deviation)
  - Once probability distribution is known, use it to estimate the conditional probability $P(X_i|Y)$

Estimate Probabilities from Data

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

- Normal distribution:

$$P(X_i \mid Y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(X_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

  – One for each (X_i, Y_i) pair

- For (Income, Class=No):
  - If Class=No
  - sample mean = 110
  - sample variance = 2975

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

Example of Naïve Bayes Classifier

Given a Test Record:

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

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

Naïve 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) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
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 | No) = P(Refund=No | No)
$\times$ P(Divorced | No)
$\times$ P(Income=120K | No)
= 4/7 $\times$ 1/7 $\times$ 0.0072 = 0.0006

P(X | Yes) = P(Refund=No | Yes)
$\times$ P(Divorced | Yes)
$\times$ P(Income=120K | Yes)
= 1 $\times$ 1/3 $\times$ 1.2 $\times 10^{-9}$ = 4 $\times 10^{-10}$

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

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

Issues with Naïve Bayes Classifier

Naïve 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) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
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(Yes) = 3/10
  P(No) = 7/10

- P(Yes | Married) =P(Married | Yes) P(Yes) / P(Married)
- =0 x 3/10 / P(Married)

P(No | Married) = P(Married | No) P(No) / P(Married)
= 4/7 x 7/10 / P(Married)

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

Issues with Naïve Bayes Classifier

Consider the table with Tid = 7 deleted

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**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Naïve Bayes Classifier:

P(Refund = Yes | No) = 2/6
P(Refund = No | No) = 4/6
P(Refund = Yes | Yes) = 0
P(Refund = No | Yes) = 1
P(Marital Status = Single | No) = 2/6
P(Marital Status = Divorced | No) = 0
P(Marital Status = Married | No) = 4/6
P(Marital Status = Single | Yes) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
P(Marital Status = Married | Yes) = 0/3

For Taxable Income:
If class = No: sample mean = 91
sample variance = 685
If class = No: sample mean = 90
sample variance = 25

Given X = (Refund = Yes, Divorced, 120K)

P(X | No) = 2/6 X 0 X 0.0083 = 0

P(X | Yes) = 0 X 1/3 X 1.2 X 10^{-9} = 0

Naïve Bayes will not be able to classify X as Yes or No!

Issues with Naïve Bayes Classifier

- If one of the conditional probabilities is zero, then the entire expression becomes zero
- Need to use other estimates of conditional probabilities than simple fractions
- Probability estimation:

$$\text{Original}: P(A_i \mid C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace}: P(A_i \mid C) = \frac{N_{ic} + 1}{N_c + c}$$

$$\text{m-estimate}: P(A_i \mid C) = \frac{N_{ic} + mp}{N_c + m}$$

c: number of classes

p: prior probability of the class

m: parameter

N_c: number of instances in the class

N_{ic}: number of instances having attribute value A_i in class c

Issues with Naïve Bayes Classifier

Consider the table with Tid = 7 deleted

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
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

For class Yes, let m=3, p=1/3.

mp=1

For class No, let m=3, p=2/3.

mp=2

Given X=(Refund=Yes, Divorced, 120k)

P(X|NO)=(2+2)/(6+3) * 2/(6+3) *…>0

P(X|Yes)=1/(3+3) * (1+1)/(3+3) *…>0

Given X = (Refund = Yes, Divorced, 120K)

P(X | No) = 2/6 X 0 X 0.0083 = 0

P(X | Yes) = 0 X 1/3 X 1.2 X 10^{-9} = 0

Example of Naïve Bayes Classifier

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

A: attributes

M: mammals

N: non-mammals

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

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

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

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

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

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

=> Mammals

Target is to predict P(M/A), P(N/A)

Example of Naïve Bayes Classifier

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

A: attributes

M: mammals

N: non-mammals

$$P(A|M) = \frac{1}{7} \times \frac{1}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.0017$$

$$P(A|N) = \frac{12}{13} \times \frac{3}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0151$$

$$P(A|M)P(M) = 0.0017 \times \frac{7}{20}$$

$$P(A|N)P(N) = 0.0151 \times \frac{13}{20}$$

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

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

=> NON-Mammals

Example 2

- Consider the training data set. There are three attributes A, B, and C. The class label is in column Y. Predict the class label for a test sample (A=1, B=0, C=0) using the naïve Bayes classifier. The answer can be +, -, or cannot decide.

Record	A	B	C	Y
1	1	0	1	-
2	0	2	0	+
3	1	1	0	+
4	0	1	1	-
5	0	0	0	-
6	0	2	1	+
7	1	1	0	-
8	1	2	1	-
9	0	2	0	+
10	1	1	1	-

$P(Y=+|A=1, B=0, C=0)= P(A=1|+)P(B=0|+)P(C=0|+)P(+)/P(1,0,0) =1/4*0*3/4*0.4=0$

$P(Y=-|A=1,B=0,C=0)=P(A=1|-)P(B=0|-)P(C=0|-)P(-)/P(1,0,0) =4/6*2/6*2/6*0.6=2/45=0.044$

Naïve Bayes (Summary)

- Bayes rule lets us do diagnostic queries with causal probabilities
- The naïve Bayes assumption takes all features to be independent given the class label
- We can build classifiers out of a naïve Bayes model using training data
- Robust to isolated noise points
- Robust to irrelevant attributes

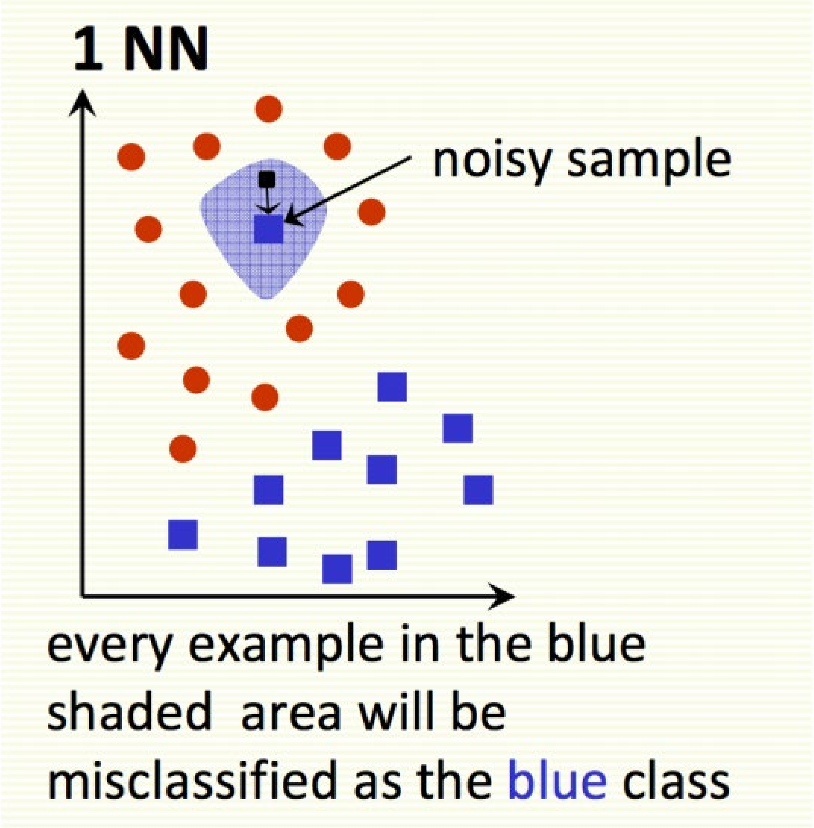
- Bayes based classifier
- KNN

Nearest neighbors

- Nearest neighbor classifier are based on learning by analogy, that is, by comparing a giving test tuple with training tuples that are similar to it.
- The training tuples are described by n attributes.
- When K=1, the unknown tuple is assigned the class of the training tuple that is closest to it in pattern space.

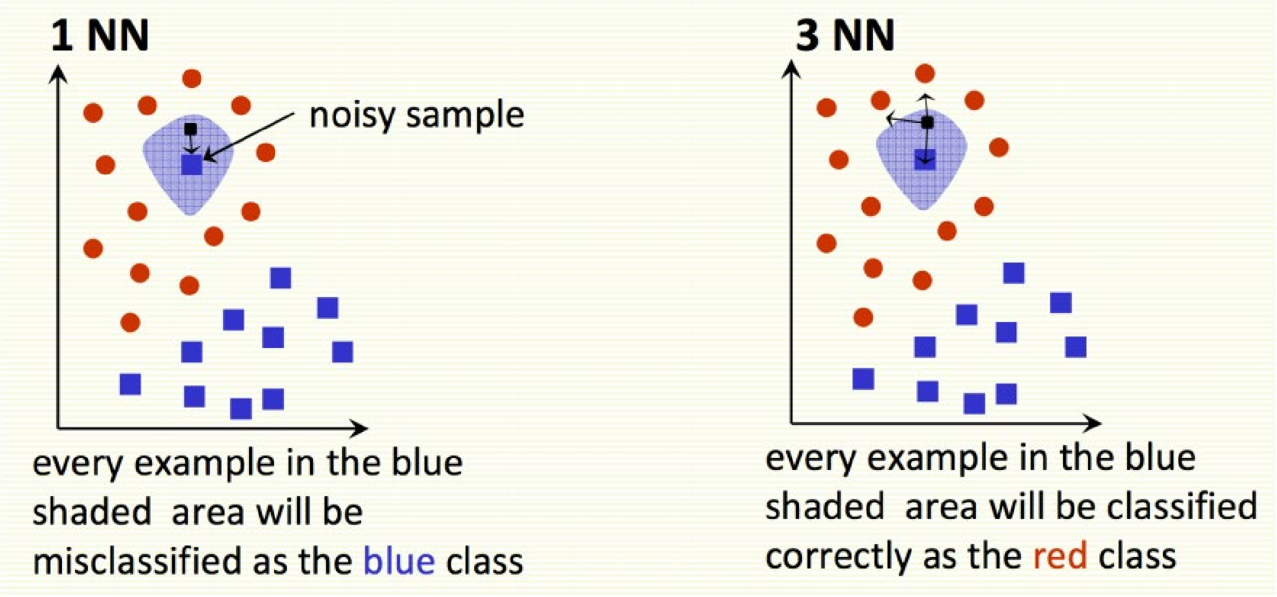
Nearest neighbors

- Nearest neighbors: sensitive to mislabeled data ("class noise").



k-Nearest Neighbors

- Smooth by having k nearest neighbors vote



- Algorithms

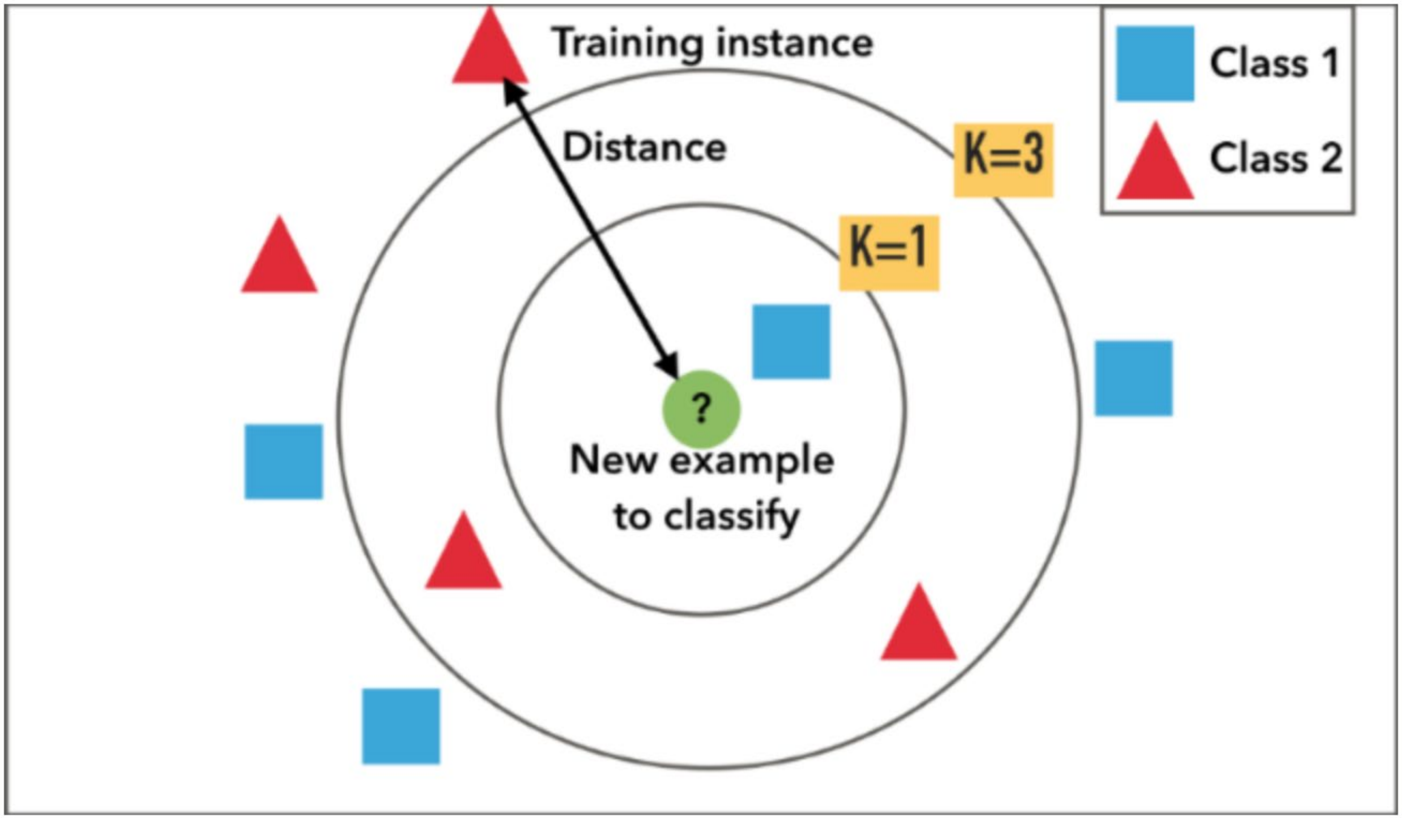
Algorithm (kNN):

1. Find k examples $\{\mathbf{x}^{(i)}, t^{(i)}\}$ closest to the test instance $\mathbf{x}$
2. Classification output is majority class

$$y = arg \max_{t^{(z)}} \sum_{r=1}^{k} \delta(t^{(z)}, t^{(r)})$$

When K=1 or 3?

- 1/3-Nearest Neighbor



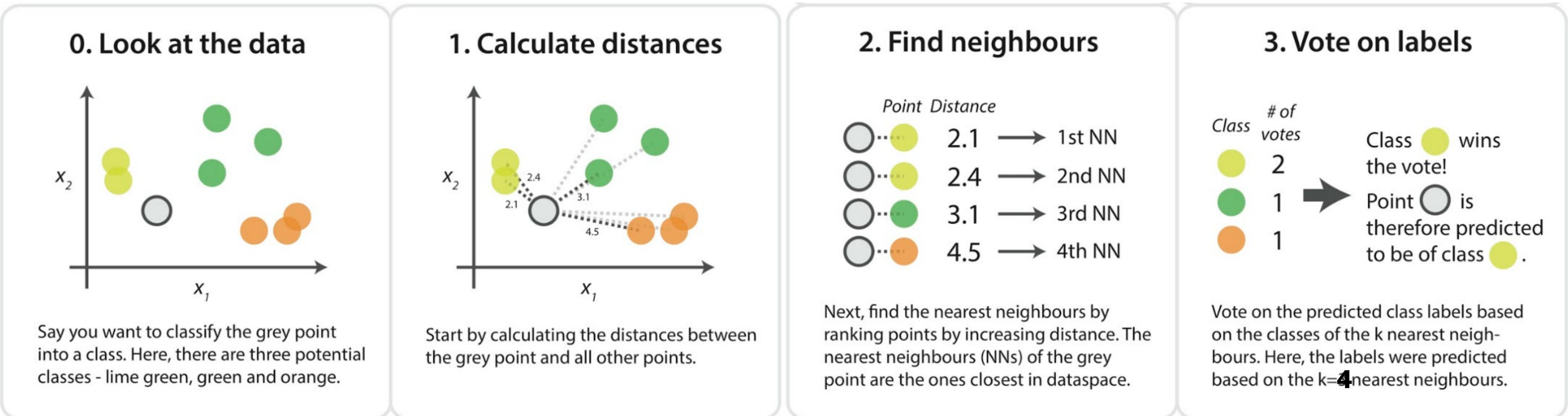
K-Nearest Neighbour algorithm

- Given a new set of measurements, perform the following steps:

  1. Pick a value for k
  2. Starting with object i,
  3. Find the k nearest objects in the training set according to euclidean distance
  4. Among these k entities, which label is most common? Pick that label for the object i.
  5. Repeat 2-5 until all i have been classified

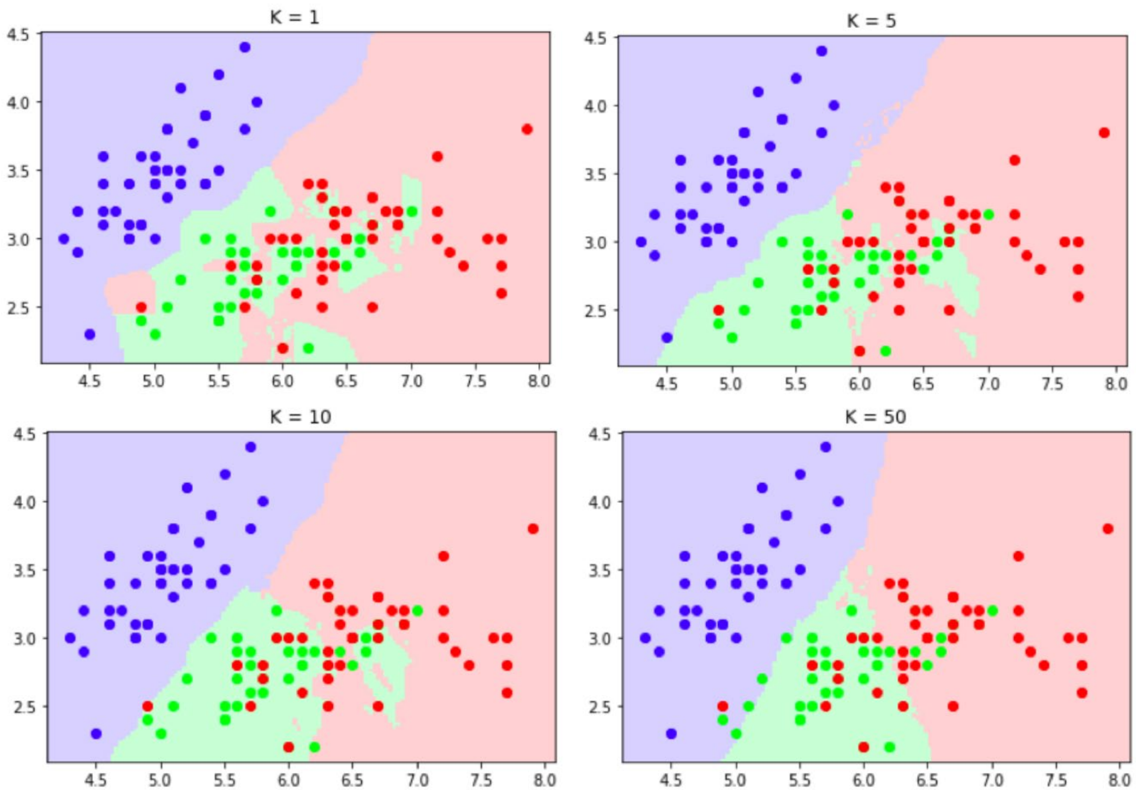
K-nearest neighbor (kNN)

- Examples with K=4.



Effect of K

- Large K yields smoother predictions, since we average over more data



Visualize the result based on different K

- We can see that when K is small, there are some outliers of green label are still green, and outliers of red label are still red.
- When K becomes larger, the boundary is more consistent and reasonable.

Nearest Neighbors: The basic version

- Training examples are vectors x_i associated with a label y_i
  - E.g. x_i = a feature vector for an email, y_i = SPAM
- Learning: Just store all the training examples
- Prediction for a new example x
  - Find the training example x_i that is closest to x
  - Predict the label of x to the label y_i associated with x_i
  - For classification: Every neighbor votes on the label. Predict the most frequent label among the neighbors.

Instance based learning

- A class of learning methods
  - Learning: Storing examples with labels
  - Prediction: When presented a new example, classify the labels using similar stored examples

- K-nearest neighbors algorithm is an example of this class of methods

- Also called lazy learning, because most of the computation (in the simplest case, all computation) is performed only at prediction time

How do we measure distances between instances?

- Numeric features, represented as n dimensional vectors
- Euclidean distance

$$||\mathbf{x}_1 - \mathbf{x}_2||_2 = \sqrt{\sum_{i=1}^{n} (\mathbf{x}_{1,i} - \mathbf{x}_{2,i})^2}$$

- Manhattan distance

$$||\mathbf{x}_1 - \mathbf{x}_2||_1 = \sum_{i=1}^{n} |\mathbf{x}_{1,i} - \mathbf{x}_{2,i}|$$

- Lp-norm
  - Euclidean = L2
  - Manhattan = L1

$$||\mathbf{x}_1 - \mathbf{x}_2||_p = \left(\sum_{i=1}^{n} |\mathbf{x}_{1,i} - \mathbf{x}_{2,i}|^p\right)^{\frac{1}{p}}$$

Distance between instances

- Symbolic/categorical features
- Most common distance is the Hamming distance
  - Number of bits that are different
  - Or: Number of features that have a different value
  - Also called the overlap
- Example:
  - X1: {Shape=Triangle, Color=Red, Location=Left, Orientation=Up}
  - X2: {Shape=Triangle, Color=Blue, Location=Left, Orientation=Down}

- Hamming distance = 2

Example 1

We have data from the questionnaires survey (to ask people opinion) and objective testing with two attributes (acid durability and strength) to classify whether a special paper tissue is good or not. Here is four training samples

X1 = Acid Durability (seconds)	X2 = Strength (kg/square meter)	Y = Classification
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good

Now the factory produces a new paper tissue that pass laboratory test with X1 = 3 and X2 = 7. Without another expensive survey, can we guess what the classification of this new tissue is?

Example 1

1. Determine parameter K = number of nearest neighbors

Suppose use K = 3

2. Calculate the distance between the query-instance and all the training samples

Coordinate of query instance is (3, 7), instead of calculating the distance we compute square distance which is faster to calculate (without square root)

X1 = Acid Durability (seconds)	X2 = Strength (kg/square meter)	**Square Distance to query instance (3, 7)**
7	7	$(7-3)^2+(7-7)^2=16$
7	4	$(7-3)^2+(4-7)^2=25$
3	4	$(3-3)^2+(4-7)^2=9$
1	4	$(1-3)^2+(4-7)^2=13$

Example 1

3. *Sort the distance and determine nearest neighbors based on the K-th minimum distance*

X1 = Acid Durability (seconds)	X2 = Strength (kg/square meter)	Square Distance to query instance (3, 7)	**Rank minimum distance**	**Is it included in 3-Nearest neighbors?**
7	7	$(7-3)^2+(7-7)^2=16$	**3**	**Yes**
7	4	$(7-3)^2+(4-7)^2=25$	**4**	**No**
3	4	$(3-3)^2+(4-7)^2=9$	**1**	**Yes**
1	4	$(1-3)^2+(4-7)^2=13$	**2**	**Yes**

Example 1

4. Gather the category Y of the nearest neighbors. Notice in the second row last column that the category of nearest neighbor (Y) is not included because the rank of this data is more than 3 (=K).

X1 = Acid Durability (seconds)	X2 = Strength (kg/square meter)	Square Distance to query instance (3, 7)	Rank minimum distance	Is it included in 3-Nearest neighbors?	**Y = Category of nearest Neighbor**
7	7	$(7-3)^2+(7-7)^2=16$	3	Yes	**Bad**
7	4	$(7-3)^2+(4-7)^2=25$	4	No	-
3	4	$(3-3)^2+(4-7)^2=9$	1	Yes	**Good**
1	4	$(1-3)^2+(4-7)^2=13$	2	Yes	**Good**

5. *Use simple majority of the category of nearest neighbors as the prediction value of the query instance*

We have 2 good and 1 bad, since 2>1 then we conclude that a new paper tissue that pass laboratory test with X1 = 3 and X2 = 7 is included in **Good** category.

Example 2

- Have the following datasets

Sepal Length	Sepal Width	Species
5.3	3.7	Setosa
5.1	3.8	Setosa
7.2	3.0	Virginica
5.4	3.4	Setosa
5.1	3.3	Setosa
5.4	3.9	Setosa
7.4	2.8	Virginica
6.1	2.8	Verscicolor
7.3	2.9	Virginica
6.0	2.7	Verscicolor
5.8	2.8	Virginica
6.3	2.3	Verscicolor
5.1	2.5	Verscicolor
6.3	2.5	Verscicolor
5.5	2.4	Verscicolor

- Please identify the new unlabeled flower

Sepal Length	Sepal Width	Species
5.2	3.1	?

Example 2

- First calculated the distances (here we just utilized Euclidean distance)

Sepal Length	Sepal Width	Species	Distance
5.3	3.7	Setosa	0.608
5.1	3.8	Setosa	0.707
7.2	3.0	Virginica	2.002
5.4	3.4	Setosa	0.36
5.1	3.3	Setosa	0.22
5.4	3.9	Setosa	0.82
7.4	2.8	Virginica	2.22
6.1	2.8	Verscicolor	0.94
7.3	2.9	Virginica	2.1
6.0	2.7	Verscicolor	0.89
5.8	2.8	Virginica	0.67
6.3	2.3	Verscicolor	1.36
5.1	2.5	Verscicolor	0.60
6.3	2.5	Verscicolor	1.25
5.5	2.4	Verscicolor	0.75

Example 2

- Find the rank

Sepal Length	Sepal Width	Species	Distance	Rank
5.3	3.7	Setosa	0.608	3
5.1	3.8	Setosa	0.707	6
7.2	3.0	Virginica	2.002	13
5.4	3.4	Setosa	0.36	2
5.1	3.3	Setosa	0.22	1
5.4	3.9	Setosa	0.82	8
7.4	2.8	Virginica	2.22	15
6.1	2.8	Verscicolor	0.94	10
7.3	2.9	Virginica	2.1	14
6.0	2.7	Verscicolor	0.89	9
5.8	2.8	Virginica	0.67	5
6.3	2.3	Verscicolor	1.36	12
5.1	2.5	Verscicolor	0.60	4
6.3	2.5	Verscicolor	1.25	11
5.5	2.4	Verscicolor	0.75	7

Example 2

- If K=1

Sepal Length	Sepal Width	Species	Distance	Rank
5.1	3.3	Setosa	0.22	1

Nearest Neighbors for 1

- If K=2

Sepal Length	Sepal Width	Species	Distance	Rank
5.1	3.3	Setosa	0.22	1
5.4	3.4	Setosa	0.36	2

Nearest Neighbors for 2

- If K=5

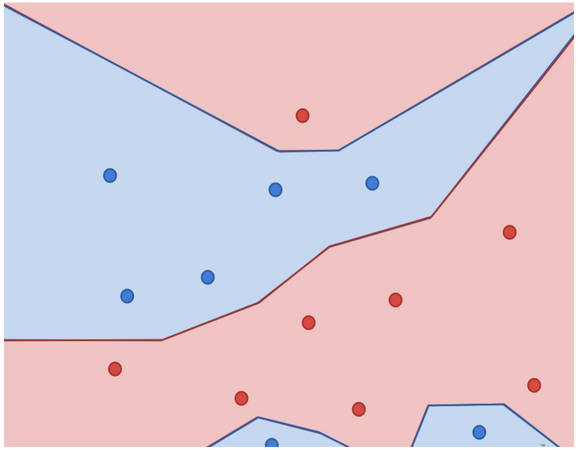
Sepal Length	Sepal Width	Species	Distance	Rank
5.1	3.3	Setosa	0.22	1
5.4	3.4	Setosa	0.36	2
5.1	3.7	Setosa	0.608	3
5.1	2.5	Verscicolor	0.6	4
5.8	2.8	Virginica	0.67	5

Nearest Neighbors for 5

Sepal Length	Sepal Width	Species	Distance	Rank
5.3	3.7	Setosa	0.608	3
5.1	3.8	Setosa	0.707	6
7.2	3.0	Virginica	2.002	13
5.4	3.4	Setosa	0.36	2
5.1	3.3	Setosa	0.22	1
5.4	3.9	Setosa	0.82	8
7.4	2.8	Virginica	2.22	15
6.1	2.8	Verscicolor	0.94	10
7.3	2.9	Virginica	2.1	14
6.0	2.7	Verscicolor	0.89	9
5.8	2.8	Virginica	0.67	5
6.3	2.3	Verscicolor	1.36	12
5.1	2.5	Verscicolor	0.60	4
6.3	2.5	Verscicolor	1.25	11
5.5	2.4	Verscicolor	0.75	7

Advantages

- Training is very fast
  - Just adding labeled instances to a list
  - More complex indexing methods can be used, which slow down learning slightly to make prediction faster



- Can learn very complex functions
- We always have the training data
  - For other learning algorithms, after training, we don't store the data anymore. What if we want to do something with it later…

Disadvantages

- Needs a lot of storage
  - Is this really a problem now?

- Prediction can be slow!
  - Naïvely: O(dN) for N training examples in d dimensions
  - More data will make it slower
  - Compare to other classifiers, where prediction is very fast

- Nearest neighbors are fooled by irrelevant attributes
  - Important and subtle

Summary: K-Nearest Neighbors

- Guarantee: If there are enough training examples, the error of the nearest neighbor classifier will converge to the error of the optimal (i.e. best possible) predictor
- In practice, use an odd K. Why?
  - To make the sole results
- How to choose K? Using a held-out set or by cross-validation
  - Feature normalization could be important
  - Often, good idea to center the features to make them zero mean and unit standard deviation.
  - Because different features could have different scales (weight, height, etc); but the distance weights them equally
- Variants exist
  - Neighbors' labels could be weighted by their distance