Predictive Model kNN

Dr Liew How Hui

Jan 2022

Review

Week 1:

- CRISP-DM (Data Mining ⇒ Original Data (Noisy) → Clean Data ↔ (Data Preprocessing →) Pred. Model → Validation → Deployment in Prediction / Inference)

Relevant Practicals:

- p03_knn1.R
- p04_knn2.R

Outline

1 kNN Models

2 More Performance Evaluation

Theory Behind kNN

The *k-nearest neighbours* (kNN) algorithm is a non-parametric method used for classification and regression.

The assumption of kNN is “similar inputs have similar outputs”. Based on this assumption, a test input $\boldsymbol{x}$ should be assigned the most common label amongst its k most similar training inputs.

Theory (cont)

Given a positive integer k and an input $\boldsymbol{x}$, the kNN algorithm first identifies the k points in the training data $(\boldsymbol{x}_i, y_i)$ that are “closest” to $\boldsymbol{x}$, represented by $N(\boldsymbol{x})$.

- For kNN classifier, the prediction is

$$h(\boldsymbol{x}) = \text{mode}(\{y'' : (\mathbf{x}'', y'') \in N(\boldsymbol{x})\}),$$

$$\mathbb{P}(Y = j | \boldsymbol{X} = \boldsymbol{x}) = \frac{1}{k} \sum_{\boldsymbol{x}_i \in N(\boldsymbol{x})} I(y_i = j).$$

- For kNN regressor, the prediction is

$$h(\boldsymbol{x}) = \frac{1}{k} \sum_{(\boldsymbol{x}'', y'') \in N(\boldsymbol{x})} y''.$$

Distance

The kNN algorithm fundamentally relies on a “distance” metric.

What is a “distance”?

A *distance* $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is a function which satisfies the following conditions: For any $\boldsymbol{x}_i$, $\boldsymbol{x}_j$, $\boldsymbol{x}_k$,

(i) $d(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0$ and $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0$ iff $\boldsymbol{x}_i = \boldsymbol{x}_j$;

(ii) $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = d(\boldsymbol{x}_j, \boldsymbol{x}_i)$; and

(iii) $d(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq d(\boldsymbol{x}_i, \boldsymbol{x}_k) + d(\boldsymbol{x}_k, \boldsymbol{x}_j)$ (triangle inequality).

Distance (cont)

The most common choice is the *Minkowski distance*:

$$d(\boldsymbol{x}, \boldsymbol{z}) = \|\boldsymbol{x} - \boldsymbol{z}\|_r = \left(\sum_{i=1}^{p} |x_i - z_i|^r \right)^{\frac{1}{r}}, \quad \boldsymbol{x},\ \boldsymbol{z} \in \mathbb{R}^p. \quad (1)$$

Note that $\|\cdot\|^r$ is called the ℓ^r norm.
When $r = 1$, we have the *Manhattan distance*:
$\|\boldsymbol{x} - \boldsymbol{z}\|_1 = |x_1 - z_1| + |x_2 - z_2| + \cdots + |x_p - z_p|$.
When $r = 2$, we have the *Euclidean distance*:
$\|\boldsymbol{x} - \boldsymbol{z}\|_2 = \sqrt{(x_1 - z_1)^2 + (x_2 - z_2)^2 + \cdots + (x_p - z_p)^2}$.

Distance (cont)

When $r = \infty$, we have the *Chebyshev distance*:
$||\boldsymbol{x} - \boldsymbol{z}||_\infty = \max\{|x_1 - z_1|, |x_2 - z_2|, \cdots, |x_p - z_p|\}$.

The Euclidean distance is more sensitive to outliers than the Manhattan distance. When outliers are rare, the Euclidean distance performs very well and is generally preferred. When the outliers are significant, the Manhattan distance is more stable.

Question

Which of the following distance metric can not be used in k-NN?
A) Gower
B) Minkowski
C) Tanimoto
D) Jaccard
E) Mahalanobis
F) All can be used

Theory: Bayes Optimal Classifier

If we **know** $\mathbb{P}(y|\boldsymbol{x})$ (which is almost never the case), then the “optimal” prediction is

$$y^* = h_{\text{opt}}(\boldsymbol{x}) = \operatorname*{argmax}_y \mathbb{P}(y|\boldsymbol{x})$$

For this type of classifier, the **error rate** is

$$\epsilon_{BayesOpt} = 1 - \mathbb{P}(y^*|\boldsymbol{x})$$

Why is the Bayes optimal classifier interesting, if it cannot be used in practice?
It provides a **lower bound of the error rate**.

Theory (cont)

What is the **upper bound on the error**?
The constant classifier.
What about the kNN Classifier?

Theorem

As $n \to \infty$, the 1-NN error is no more than twice the error of the Bayes Optimal classifier:

$$\epsilon_{\mathrm{BayesOpt}} \leq \epsilon_{NN} \leq 2\epsilon_{\mathrm{BayesOpt}}$$

Similar guarantees hold for $k > 1$.

Implementation of kNN Classifier

A naive Python + Numpy implementation is listed below.

```
# https://github.com/joelgrus/data-science-from-scratch
from scipy.spatial.distance import euclidean
import logging
logging.basicConfig(level=logging.DEBUG)

def knn_classify(k: int, X, y, x_new, distance=euclidean) -> str:
    import numpy as np, collections
    dist_list = [distance(x, x_new) for x in X]
    logging.debug(’distances={}’.format([round(d,4) for d in dist_list]))
    order = np.argsort(dist_list)
    k_nearest_labels = y[order[:k]]
    logging.debug(’knn={}’.format(k_nearest_labels))
    vote_counts = collections.Counter(k_nearest_labels)
    winner, winner_count = vote_counts.most_common(1)[0]
    return winner, winner_count/k

if __name__==’__main__’:
    from sklearn import datasets
    iris_df = datasets.load_iris()
    X = iris_df[’data’]; y = iris_df[’target’]
    knn_classify(1,X,y,[6,3,5,2]); knn_classify(1,X,y,[5,3,1,0])
```

Example of kNN Classifier

A sport school would like to group their new enrolled students into 2 groups, as according to the existing students' weight and height. The weight and height of 7 existing students with group are shown in the table below.

Student	Weight (kg)	Height (cm)	Group
A	29	118	A
B	53	137	B
C	38	127	B
D	49	135	B
E	28	111	A
F	24	111	A
G	30	121	A

Example (cont)

(a) Perform and use $k = 3$-NN method (with Euclidean distance) to predict which group the following students will be grouped into, based on **cut-off of 0.7** on group A.

Student	Weight (kg)	Height (cm)
H	35	120
I	47	131
J	22	115
K	38	119
L	31	136

Example (cont)

(b) The actual groups of the students are {A, B, A, B, B} for students {H, I, J, K, L} respectively. Construct a confusion matrix and calculate the accuracy measurements for a cut-off of 0.5 and a cut-off of 0.7.

(c) Write a Python script to produce the above calculations using the simple implementation by lecturer above and also use sklearn's implementation.

(d) Is R script easier to write???

kNN Regressor and Example

The *kNN regressor* applies “mean” instead of “mode” to “predict” the output.

R: `FNN::knn.reg`

Python: `sklearn.neighbors.KNeighborsRegressor`

Example 2.4.1 (Exam SRM Study Manual, p225, Q15.10)

A continuous variable Y is modelled as a function of X using kNN with $k = 3$. With the following data:

X	5	8	15	22	30
Y	4	1	10	16	30

Calculate the fitted value of Y at $X = 12$. Try write a script using R (or Python) to perform the calculation for you.

Classifier Boundary

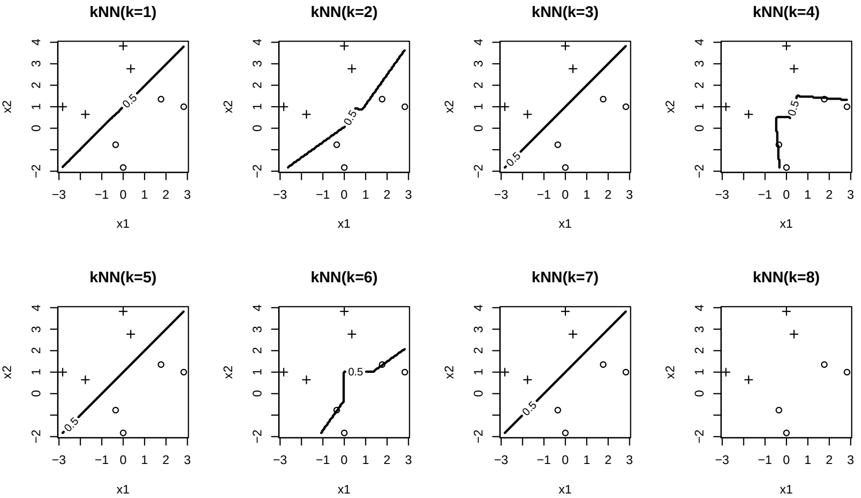
What is “classifier boundary”? It is the boundary between different classes. We can think of it as the “boundaries” between “countries”, for example, Malaysia has boundaries with Singapore, Thailand and Indonesia.

Most data are not 2D, for example, the iris data. But most analysis will try to reduce to 2D to “visualise” the boundary. We can find a comprehensive analysis of the iris data using R with beautiful diagrams on `https://www.datacamp.com/community/tutorials/machine-learning-in-r`. As for the drawing the kNN boundary between classes, a “contour” plot is required and is illustrated by

`https://stats.stackexchange.com/questions/21572/how-to-plot-decision-boundary-of-a-k-nearest-neighbor-c`

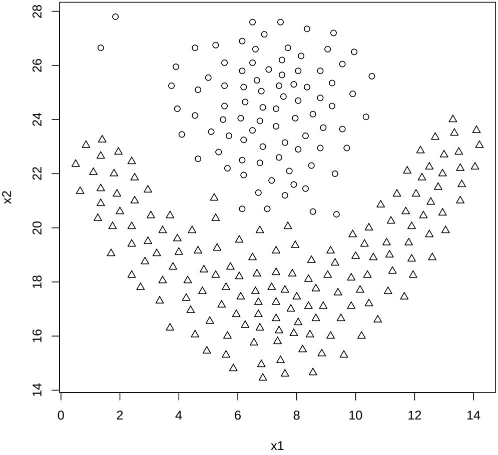
Classifier Boundary (cont)

First, consider the boundary of the ‘data1’ found in your Practical 1 which is symmetrical. Therefore, we “feel” that it is a “line”.

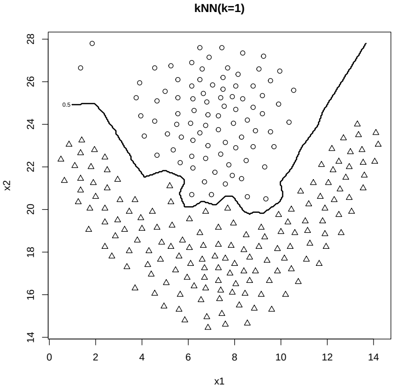


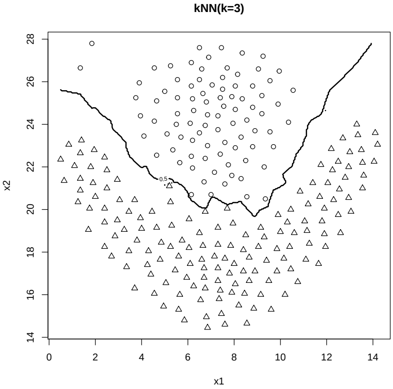
Classifier Boundary (cont)

Now, we investigate a “nonlinear boundary” data.

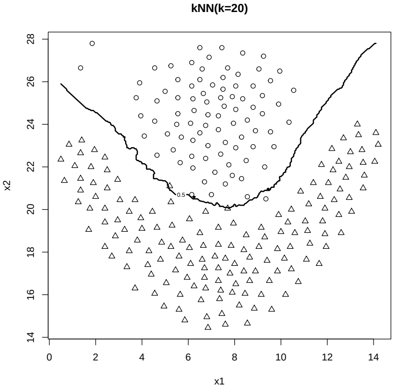


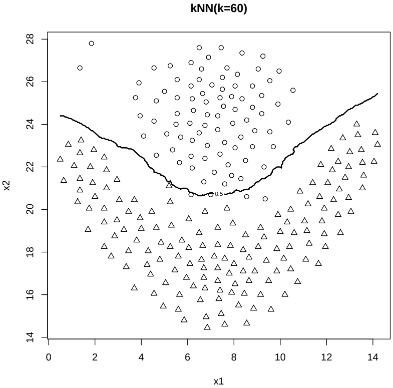
Classifier Boundary (cont)



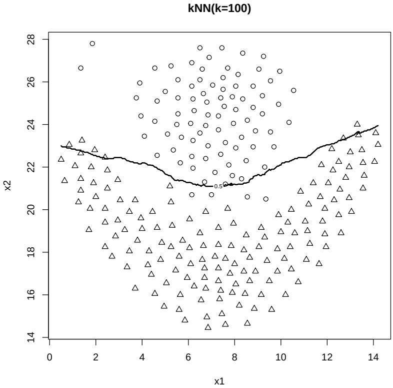


Classifier Boundary (cont)


kNN(k=20)


kNN(k=60)

Classifier Boundary (cont)



Conclusion: The choice of k depends upon the data.
In general,

- k small, decision boundary is over "flexible" ⇒ kNN classifier is low bias, high variance.
- k large, decision boundary is less flexible, "less noise", "smoother", more "linear" ⇒ kNN classifier is low-variance, high bias.

Feature Scaling / Standardisation

In the above example, if we change student weight to *g* and height to metre, what would happen?

Student	Weight (g)	Height (m)	Group
A	29,000	1.18	A
B	53,000	1.37	B
C	38,000	1.27	B
D	49,000	1.35	B
E	28,000	1.11	A
F	24,000	1.11	A
G	30,000	1.21	A

Some numbers are too large, while others are too small!

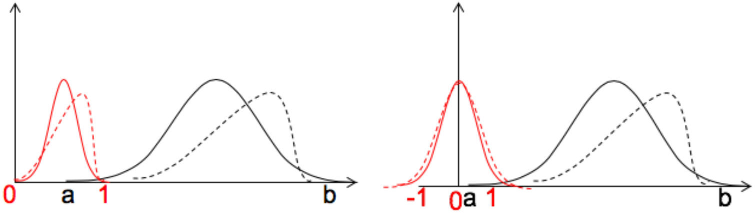
Feature Scaling (cont)

Bad for “predictive model” training.
Feature scaling / Standardisation: Put all variables into the “similar” range, the variables are equally weight.
Two famous methods:

- Min-Max Normalisation (Rescaling):
  $M(X_{ij}) = \frac{X_{ij} - X_{\text{min,j}}}{X_{\text{max,j}} - X_{\text{min,j}}}$
- Standard Scaler / Standardisation:
  $M(X_{ij}) = \frac{X_{ij} - \overline{X}_j}{s_{X,j}}$

Feature Scaling (cont)

Diagrams to illustrate the effect of min-max scaling and standardisation are



The former transforms to the range $[0, 1]$ while the later transforms to a “standard normal distributed” range.

Feature Scaling Example

Given the training data and testing data in Example of kNN Classifier, apply the standardisation to perform and use $k = 3$-NN method (with Euclidean distance) to predict which group the following students will be grouped into, based on **cut-off of 0.7** on group A.

WARNING

In Final Exam / Assessment, if the question did not mention “feature scaling” you DON’T NEED TO SCALE the features.

In real-world situation, you want perform feature scaling a lot of the times using computer.

Feature Scaling Example (cont)

Student	Weight (kg)	Height	Norm_Wgt	Norm_Hgt	Group
A	29	118	-0.6113	-0.4587	A
B	53	137	1.5284	1.3355	B
C	38	127	0.1910	0.3912	B
D	49	135	1.1717	1.1467	B
E	28	111	-0.7005	-1.1197	A
F	24	111	-1.0571	-1.1197	A
G	30	121	-0.5222	-0.1754	A
Mean	35.8571	122.8571			
Std. dev	11.2165	10.5898			

Feature Scaling Example (cont)

$k = 3$, cut-off = 0.7:

		H	I	J	K	L
	Group	Distance	Distance	Distance	Distance	Distance
A	A	0.5673	2.0205	0.6854	0.8079	1.7091
B	B	2.2699	0.7792	3.4575	2.1628	1.9637
C	B	0.7131	0.8869	1.8218	0.7554	1.0544
D	B	1.8879	0.4177	3.0596	1.8013	1.6076
E	A	1.0544	2.5370	0.6548	1.1686	2.3759
F	A	1.2977	2.7878	0.4177	1.4590	2.4419
G	A	0.4557	1.7857	0.9109	0.7378	1.4193
$P(Y = A)$		0.6667	0.0000	1.0000	0.6667	0.3333
Cut-off = 0.5	$\hat{y}$	A	B	A	A	B
Cut-off = 0.7	$\hat{y}$	B	B	A	B	B

Weighted kNN

class::knn → kNN classifier with Euclidean distance
FNN::knn.reg → kNN regressor with Euclidean distance
kknn::kknn → Weighted kNN (wkNN)

1. Let $N(\boldsymbol{x}, k+1)$ be the $k+1$ nearest neighbours to $\boldsymbol{x}$ according to a distance function $d(\boldsymbol{x}, \boldsymbol{x}_i)$.
2. The $(k+1)$th neighbour is used for "standardisation" of the k smallest distances via $D_i = \frac{d(\boldsymbol{x},\boldsymbol{x}_i)}{d(\boldsymbol{x},\boldsymbol{x}_{k+1})}$.
3. A weighted majority of the k nearest neighbour

$$\hat{y} = \max_j \left\{ \sum_{i=1}^{k} K(D_i) I(y_i = j) \right\}.$$

Weighted kNN (cont)

K is called the *kernel (function)* if it satisfies

1. $K(x) \geq 0$ for all $x \in \mathbb{R}$;
2. $K(x)$ is maximum when $x = 0$;
3. $K(x)$ descents monotonously when $x \to \pm\infty$.

A *rectangular kernel*, $K(x) = \frac{1}{2}I(|x| \leq 1)$; A *triangular kernel*, $K(x) = (1 - |x|) \cdot I(|x| \leq 1)$.

wkNN = kNN when kernel=rectangular (R's kknn) or weights=uniform (Python)

Outline

1 kNN Models

2 More Performance Evaluation

Recall: Data Science and CRISP-DM

CRISP-DM (Cross Industry Standard Process for Data Mining)

- Business understanding
- Data understanding
- Data preparation / preprocessing
- Modelling
- Evaluation or Validation
- Deployment

The validation methods we learned in Week 1 is the **holdout/validation set** method. For regression problems, the measures are RSS(Square Error), R^2; For classification problem, the measures are accuracy; in particular, for binary classification problem, confusion matrix / contingency table, ROC.

More Performance Evaluation

There are a few things we need to take note regarding the holdout/validation set method:

- It works only for time-independent data. For a time-related data (e.g. the ISLR's Smarket data in the Practical 3 script `p03_knn1.R`), we have to split the data by earlier period and later period based on some cut-off time.
- For classification problem, linear sampling is allowed when the size of the data is large and the classes are 'uniformly' distributed, otherwise, stratified sampling should be used.

More Performance Evaluation (cont)

The holdout/validation set method (cont):

- The result depends on the ‘selection’ of data ‘randomly’. Note that in Practical 3, there is a `set.seed(123)` which fixed the ‘selection’. If we change the number 123 to 124 or 125, we will get different results for the performance. If we apply stratified sampling with odd index, we get an accuracy of 0.8011988 and with even index, we get an accuracy of 0.7654691

This point leads to the question: What is the ‘real’ performance of the (kNN) model?

More Performance Evaluation (cont)

For a data with 1000 rows, if we choose 700 rows for training and 300 rows for testing using linear sampling, there are

$$\frac{1000!}{700!300!}$$

possible combinations! The number is just **too large**.

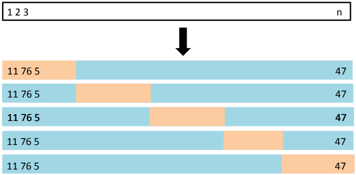
We can try multiple samples to have a better confidence on the performance for the holdout method or we can use the k-fold cross-validation method mentioned next.

Cross-Validation (cont)

k ***-fold cross validation (CV)***

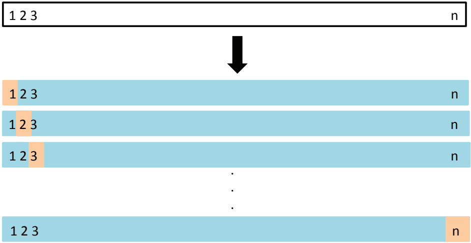
- Randomly divides the set of observations into *k* “equal” “folds”;
- First fold = validation set & remaining $k - 1$ folds = training set.
- Second fold = validation set & remaining $k - 1$ folds = training set.
- etc.

5-fold example:



Cross-Validation (cont)

Special case $k = n \Rightarrow$ **leave-one-out cross validation (LOOCV)**



Since training sample is close to n, the bias is small. The variance is usually high because n training sets are similar to each other due to strong correlation.

Cross-Validation (cont)

Software support:
R: `caret::createDataPartition`, `createResamples` (for bootstrapping), `createFolds` (for *k*-fold cross validation), `createMultiFolds` (for repeated cross-validation), `createTimeSlices` (for dealing with time-series).
Python: `train_test_split` (simple random split), `KFold` or `StratifiedKFold` (*k*-fold CV), `LeaveOneOut` (LOOCV) from `sklearn.cross_validation`.

Cross-Validation (cont)

If we don’t have the right to install additional package like the caret library, we can still use R’s basic commands to perform k-fold cross validation. The $k = 10$ example is shown below.

https://stats.stackexchange.com/questions/61090/how-to-split-a-data-set-to-do-10-fold-cross-validation

```
#Randomly shuffle the data & create 10 fold
d.f = d.f[sample(nrow(d.f)),]
folds = cut(seq(1,nrow(d.f)),breaks=10,labels=FALSE)
#Perform 10 fold cross validation
for(i in 1:10){
    #Segement your data by fold using which()
    testIndexes = which(folds==i,arr.ind=TRUE)
    testData  = d.f[testIndexes, ]
    trainData = d.f[-testIndexes, ]
    ...
}
```

Cross-Validation (cont)

Using `caret` library (it is a complex R packages with many dependencies):

```
set.seed(123)
train.control <- trainControl(method="cv", number=5)
model <- train(y ~., data=train.data, method="lm",
               trControl=train.control)
print(model)
```

When we split data into two groups — ‘training’ and ‘testing’, we will have **training errors** and **testing errors**. We are usually concerned about the testing errors but for inflexible models (like the logistic regression from next topic), we also need to look at the training errors.

Cross-Validation (cont)

Final Exam May 2019, Q3

(a) Supervised learning includes classification and regression.

    (i) State the difference between classification and regression in term of response variable. (1 mark)

    (ii) Explain the sampling methods used in splitting data for classification and regression respectively. (4 marks)

(b)

    (i) State an issue that comes along with split validation, which can be overcome by using cross validation. (1 mark)

    (ii) Describe the process of a 5-fold cross validation. (4 marks)

Cross-Validation (cont)

Final Exam May 2019, Q3 (cont)

(c) A sample of 500 males and 800 females had been collected to test on a model of gender prediction. The model resulted that 380 males and 510 females were predicted correctly.

   (i) Assume male as positive class and female as negative class, calculate the count of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) for the model's result. (2 marks)

   (ii) Construct the confusion matrix for the model. State the classification error, specificity and sensitivity of the model. (4 marks)

   (iii) Compare the recall and precision for both male and female. Interpret your results. (4 marks)