Tut 2: kNN

Jan 2023

kNN Predictive Models

kNN is discriminative, non-parametric predictive model

- For kNN classifier, the mathematical formulation is

$$\hat{h}(\boldsymbol{x}) = \underset{j\in\{1,\cdots,K\}}{\operatorname{argmax}} \frac{1}{k} \sum_{\boldsymbol{x}_i \in N(\boldsymbol{x})} I(y_i = j)$$

- For kNN regressor, the mathematical formulation is

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

One popular choice of distance in kNN is the *Minkowski distance*:

$$d(\boldsymbol{x}, \boldsymbol{z}) = \|\boldsymbol{x} - \boldsymbol{z}\|_r = \left(\sum_{i=1}^{p} |x_i - z_i|^r \right)^{\frac{1}{r}}, \quad \boldsymbol{x},\ \boldsymbol{z} \in \mathbb{R}^p. \tag{2.1}$$

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*:

$$\|\boldsymbol{x} - \boldsymbol{z}\|_2 = \sqrt{(x_1 - z_1)^2 + (x_2 - z_2)^2 + \cdots + (x_p - z_p)^2}.$$

There are other distance / dissimilarity functions which are used in specific cases:

- Gower
- Tanimoto
- Jaccard
- Mahalanobis

1. The given table provides a training data set containing six observations, three predictors and one qualitative response variable. Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using k-nearest neighbours.

Obs.	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

(a) Compute the Euclidean distance between each observation and the test point (TP).

Solution.

Obs.	X_1	X_2	X_3	Y	Distance
1	0	3	0	Red	3
2	2	0	0	Red	2
3	0	1	3	Red	$\sqrt{10} \approx 3.1623$
4	0	1	2	Green	$\sqrt{5} \approx 2.2361$
5	-1	0	1	Green	$\sqrt{2} \approx 1.4142$
6	1	1	1	Red	$\sqrt{3} \approx 1.7321$

□

(b) What is our prediction with $k = 1$? Why?

Solution. Green. Observation 5 is the closest neighbour for $k = 1$. □

(c) What is our prediction with $k = 3$? Why?

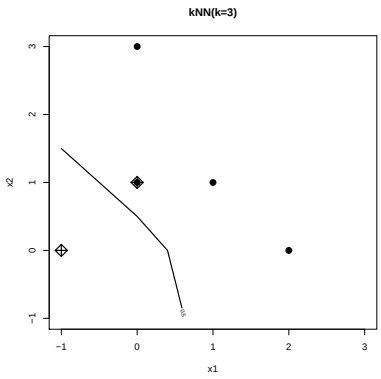
Solution. Red. Observations 2, 5 and 6 are the closest neighbours for $k = 3$, which Y equal to (Red, Green, Red). The probability of Red is two-third, which is larger than 0.5. $\boxed{\mathbb{P}(Y = \text{Red}) = \frac{2}{3} \geq 0.5}$ Hence, the test point will be predicted to be Red. □

(d) If the Bayes decision boundary in this problem is highly non-linear, then would we expect the optimum value for k to be large or small? Why?

Solution. Small. A small k would be flexible for a non-linear decision boundary, whereas a large k would try to fit a more linear boundary because it takes more points into consideration. □

(e) By considering X_1 and X_2 only, sketch the 3-nearest neighbours decision boundary for range $-1 \leq X_1 \leq 3$ and $-1 \leq X_2 \leq 3$, with the distance measure used in (a). Assume that X_1 and X_2 can only take integer values.

Solution.




```
library(class)  # for knn
d.train = data.frame(
  x1 = c(0,2,0,0,-1,1), x2 = c(3,0,1,1,0,1),
  x3 = c(0,0,3,2,1,1),
y = c('R','R','R','G','G','R'))
g.x1 = g.x2 = seq(-1,3,by=0.1)
# x1, x2 can only take integer values
d.grid = expand.grid(x1=g.x1, x2=g.x2)
predicted = knn(d.train[,-c(3,4)],test=d.grid,
  cl=d.train[,4],k=3,prob=TRUE)
prob = attr(predicted, "prob")
prob = ifelse(predicted=="G", prob, 1-prob)
prob = matrix(prob, length(g.x1), length(g.x2))
#https://stats.stackexchange.com/questions/21572/how-to-plot-decisio
contour(g.x1, g.x2, prob, levels=0.5, main="kNN(k=3)",
  xlab="x1",ylab="x2",lw=2)
attach(d.train)
points(x1, x2, pch=ifelse(y=="G",9,20),cex=2,lwd=2)
points(d.grid, col=ifelse(prob<0.5,"red","green"),cex=0.7)
```

□

More Performance Evaluation

2. (Jan 2022 Final Q1(d)) Explain the steps in (i) validation set approach and (ii) k-fold cross validation and state each advantages and disadvantages. (5 marks)

Solution. (i) Validation set approach shuffles the data and splits it into training set and validation/test set. The training dataset is the sample of data used to fit the model; the validation dataset is the sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters. The evaluation becomes more biased as skew on the validation dataset is incorporated into the model configuration. The validation dataset may also play a role in other forms of model preparation, such as feature selection. . . [1 mark]

The advantage of this approach is its simplicity in scoring predictive models. . . . . [1 mark]

The disadvantage of this approach is its biasness and being too dependent on a particular sampling. . . . . [0.5 mark]

(ii) k-fold cross validation shuffles the data and splits it into k groups. Each group will be set as validation/test set and the remainder will be set as training set and be used to score the predictiove models. . . . . [1 mark]

Advantage: This approach is less prone to biasness problem and less dependent on a particular sampling. It is used to tune model hyperparameters instead of a separate validation dataset. [1 mark]

Disadvantage: When the data size is large and the k is large, the scoring process can be very time consuming. . . . . [0.5 mark] □

3. What are the advantages of k-fold cross validation relative to

   (a) Validation set approach

   Solution. The estimate of the test error rate can be highly variable depending on which observations are included in the training and validation sets.

   Secondly, the validation set error rate may tend to overestimate the test error rate for the model fit on the entire data set, which is the overfitting problem. □

   (b) Leave-one-out cross validation (LOOCV)

Solution. LOOCV is a special case of k-fold cross-validation with $k = n$. Thus, LOOCV is the most computationally intense method since the model must be fit n times. In addition, LOOCV has **higher variance**, but **lower bias**, than k-fold cross validation. □

4. (May 2019 Final Q3)

   (a) Supervised learning includes classification and regression.

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

   Solution. The response variable for classification is categorical while for regression is numerical. □

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

   Solution. The **classification uses stratified sampling** (sklearn's `StratifiedShuffleSplit`). Samples are distributed to different sets according to the proportion of response variable. The **regression uses linear sampling** (sklearn's `ShuffleSplit`). Samples are distributed randomly to different sets. □

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

   Solution. Overfitting. □

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

   Solution. 5-fold cross validation *randomly sampled observations into 5 non-overlapping groups with equal size*, known as folds. For first iteration, first fold will be treated as validation set, and the remaining four folds act as training set. Five iterations will be run with a different fold is treated as validation set at each iteration, while the other folds served as training set. This will eventually give five estimates of accuracy measures and average will be taken. □

   (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)

   Solution. Given TP = 380; TN = 510

   $$\text{FN} = 500 - \text{TP} = 120; \quad \text{FP} = 800 - \text{TN} = 290.$$

   □

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

   Solution. Confusion matrix:

	True +	True -	Precision
Predicted +	380	290	0.5672
Predicted -	120	510	0.8095
Recall	0.7600	0.6375	0.6846

   Classification error $= 1 - 0.6846 = 0.3154$

Specificity = negative recall = 0.6375
Sensitivity = positive recall = 0.76 □

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

Solution. Male (positive class): Recall = 0.7600; Precision = 0.5672.
High recall low precision. The model is wide but generalised in predicting male. This means that the model can capture most males but those predicted males might be incorrect.
Female (negative class): Recall = 0.6375; Precision = 0.8095.
Low recall high precision. The model is small but highly specialised in predicting female. This means that the model can only capture some females but those predicted females are mostly correct. □