机器学习引论

彭玺
pengxi@scu.edu.cn
www.pengxi.me

提纲

一．Review

二．Classification Like Human – KNN classifier

三．Performance Metric

四．Model Selection and Significant Test

五．Normalization

提纲

一．Review

二．Classification Like Human – KNN classifier

三．Performance Metric

四．Model Selection and Significant Test

五．Normalization

一、Review

Vector Space

A vector space is any set *V* for which two operations are defined:

- **Vector addition**: any vector x_1 and x_2 in set *V* can be added to another vector $x = x_1 + x_2$, and their sum *x* is also in set *V*.
- **Scalar Multiplication**: Any vector *x* in *V* can be multiplied ("scaled") by a real number *c*, to produce a second vector *cx* which is also in V.

Examples: coordinate space, infinite coordinate space, Cartesian product of vector spaces, polynomial vector spaces, functional space…

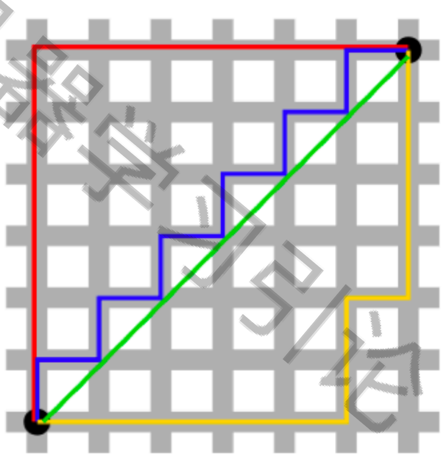
一、Review

线性空间（vector space）：

- A vector space over a field F is a set V together with **two operations** that satisfy the **eight axioms** listed below.
- The first operation, called **vector addition** or simply addition: **V + V → V**, takes any two vectors v and w and assigns to them a third vector which is commonly written as v + w, and called the sum of these two vectors. (Note that the resultant vector is also an element of the set V).
- The second operation, called **scalar multiplication**: **F × V → V**， takes any scalar a and any vector v and gives another vector av. (Similarly, the vector av is an element of the set V).
- Consists of null space (0).

一、Review

Vector Norm



Norm provides a fundamental definition of “distance” in a vector space

- 1-norm: Manhattan distance
- 2-norm: Euclidean distance (most popular, e.g., MSE, Least Squares…)
- Any other alternative?

How to measure distance between vectors?

– Obvious answer: the distance between two vectors x and y is $||x - y||$, where $||\cdot||$ is some vector norm.

– Alternative: use the angle between two vectors x and y to measure the distance between them.

– How to calculate the angle between two vectors?

Linear Independence

- Given a set of vectors $\{v_1, v_2, \cdots, n_n\} \in \mathbb{R}^m$, with $m \geq n$, consider the set of linear combinations $y = \sum_{j=1}^{n} \alpha_j v_j$ for arbitrary coefficients α_j's.
- The vectors $\{v_1, v_2, \cdots, n_n\}$ are linearly independent, if $\sum_{j=1}^{n} \alpha_j v_j = 0$, if and only if $\alpha_j = 0$ for all $j = 1, \cdots, n$.
- A set of m linearly independent vectors of $\mathbb{R}^m$ is called a basis in $\mathbb{R}^m$: any vector in $\mathbb{R}^m$ can be expressed as a linear combination of the basis vectors.

Linear independence could be an effective metric to measure the similarity/distance between two data points lying on different/same **subspaces**.

一、Review

Matrix Rank

- The rank of a matrix is the maximum number of linearly independent column vectors.
- A square matrix $A \in \mathbb{R}^{n\times n}$ with rank n is called nonsingular.
- A nonsingular matrix A has an inverse A^{-1} satisfying

$$AA^{-1} = A^{-1}A = I_n.$$

- What is the rank of an out-product matrix $xy^T \in \mathbb{R}^{m\times n}$ with $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$?
- Let $A \in \mathbb{R}^{n\times n}$ be nonsingular, and let $B = A + uv^T$ with $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^n$. Then, $B^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1+v^TA^{-1}u}$.

Q: what can be used by the matrix rank?

一、Review

Eigenvalues and eigenvectors

- Let A be a $n \times n$ matrix. The vector $v \neq 0$ that satisfies

$$Av = \lambda v$$

  for some scalar λ is called the eigenvector of A and λ is the eigenvalue corresponding to the eigenvector v.

- An example: $A = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}$

$$Av = \lambda v \to (A - \lambda I_n)v = 0 \to |A - \lambda I_n| = 0 \to \left| \begin{pmatrix} 2-\lambda & 1 \\ 1 & 3-\lambda \end{pmatrix} \right| = 0$$

  Two eigenvalues $\lambda_1 = 3.62$ and $\lambda_2 = 1.38$. and two eigenvectors:

$$v_1 = \begin{pmatrix} 0.52 \\ 0.85 \end{pmatrix}, \quad v_2 = \begin{pmatrix} 0.85 \\ -0.52 \end{pmatrix}$$

一、Review

Matrix norms

- $||A||_2 = \left(\max_i \lambda_i(A^TA)\right)^{1/2}$: square root of the largest eigenvalue of A^TA.
- $||A||_1 = \max_j \sum_{i=1}^m |a_{ij}|$: maximum over columns.
- $||A||_\infty = \max_i \sum_{j=1}^n |a_{ij}|$: maximum over rows.
- Frobenius norm: does not correspond to any vector norm.

$$||A||_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2}$$

  - Define $\text{trace}(B) = \sum_{i=1}^n b_{ii}$ for any matrix $B = (b_{ij}) \in \mathbb{R}^{n\times n}$.
  - Show that $||A||_F^2 = \text{trace}(AA^T)$.

Singular Value Decomposition (SVD)

Compute the norm of the matrix A:

$$||A||_2 = \sigma_1, \quad ||A||_F = \sqrt{\sum_{i=1}^{n} \sigma_i^2}.$$

The trace norm (or nuclear norm) of the matrix A is defined as:

$$||A||_* = \sum_{i=1}^{n} \sigma_i.$$

The trace norm has become very popular in recent years for matrix completion.

* E. J. Candés and T. Tao. The power of convex relaxation: Near-optimal matrix completion. IEEE Trans. Inform. Theory, 56(5), 2053-2080.
* E. J. Candés and B. Recht. Exact matrix completion via convex optimization. Found. of Comput. Math., 9 717-772.

Tips: A Machine Learning Method

Mathematical notation: objects in the physical world. Linear Algebra.

Objective function: relation among the objects.

Optimization: Solving the objective.

Performance metric: to evaluate the performance of the machine learning method.

提纲

一．Review

二．Classification Like Human – KNN classifier

三．Performance Metric

四．Model Selection and Significant Test

五．Normalization

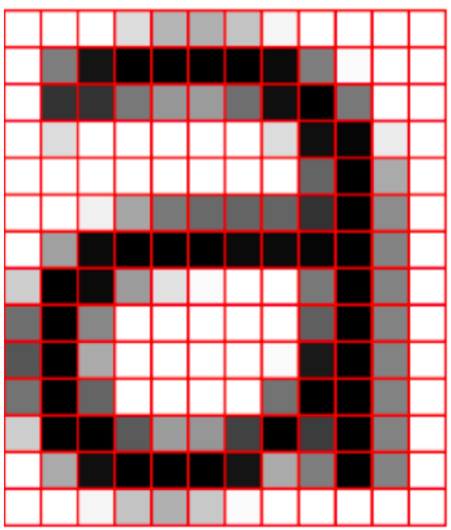
Test Questions

- Q1: What is the classification? How to perform classification by human? And what is the simplest way?
- Q2: What problem of 1NN is addressed by kNN?
- Q3: How to (why) incorporate the distance into classical kNN? And what will be benefited from it?
- Q4: How to solve the scaling issue faced by KNN?
- Q5: How to evaluate the performance of a classifier?
- Q6: What is model selection? How to solve this issue?

二、Classification Like Human

The definition of the classification?

- For a given set of two-tuple (X, Y), namely training data, one could use it to classify an unknown sample x (testing data point) based on the similarity with (X, Y), where X denotes the data point and Y is the corresponding label.



An image

1.0	1.0	1.0	0.9	0.6	0.6	0.6	1.0	1.0	1.0	1.0	1.0
1.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0	1.0	1.0
1.0	0.2	0.2	0.5	0.6	0.6	0.5	0.0	0.0	0.5	1.0	1.0
1.0	0.9	1.0	1.0	1.0	1.0	1.0	0.9	0.0	0.0	0.9	1.0
1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0
1.0	1.0	1.0	0.5	0.5	0.5	0.5	0.5	0.4	0.0	0.5	1.0
1.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0
0.9	0.0	0.0	0.6	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0
0.5	0.0	0.6	1.0	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0
0.5	0.0	0.7	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.5	1.0
0.6	0.0	0.6	1.0	1.0	1.0	1.0	0.5	0.0	0.0	0.5	1.0
0.9	0.1	0.0	0.6	0.7	0.7	0.5	0.0	0.5	0.0	0.5	1.0
1.0	0.7	0.1	0.0	0.0	0.0	0.1	0.9	0.8	0.0	0.5	1.0
1.0	1.0	1.0	0.8	0.8	0.9	1.0	1.0	1.0	1.0	1.0	1.0

A matrix/vector denoted by X_1

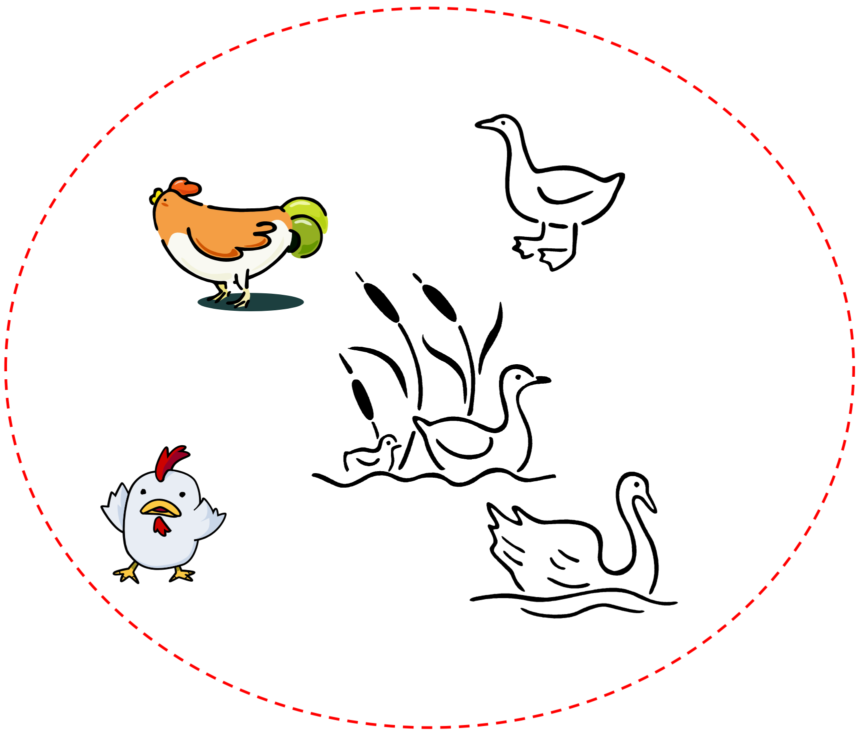
$Y_1 = 1$

Label/annotation

二、Classification Like Human

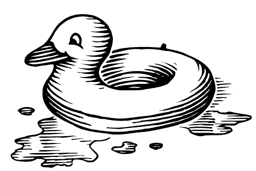
How to recognize a new object our brain?

Brain



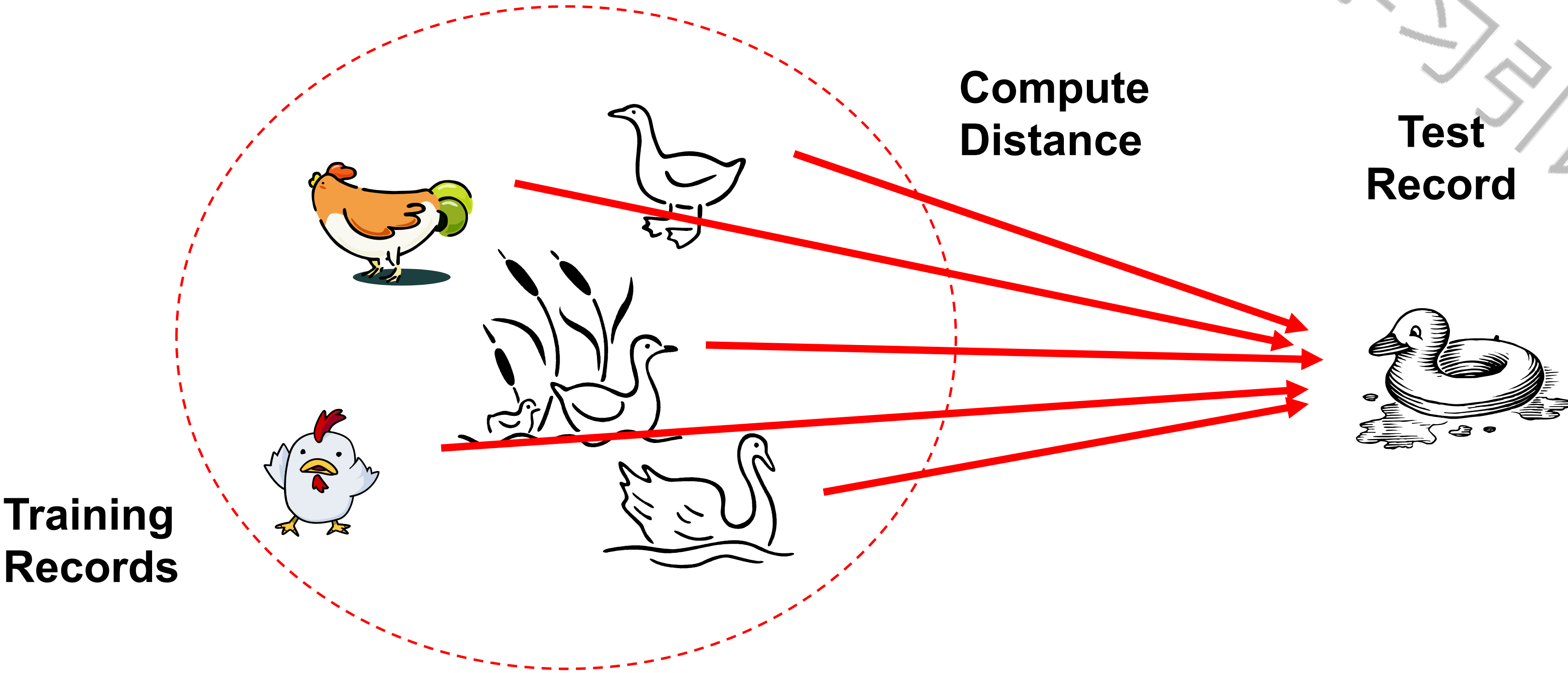
Training data -> memory

New Observation



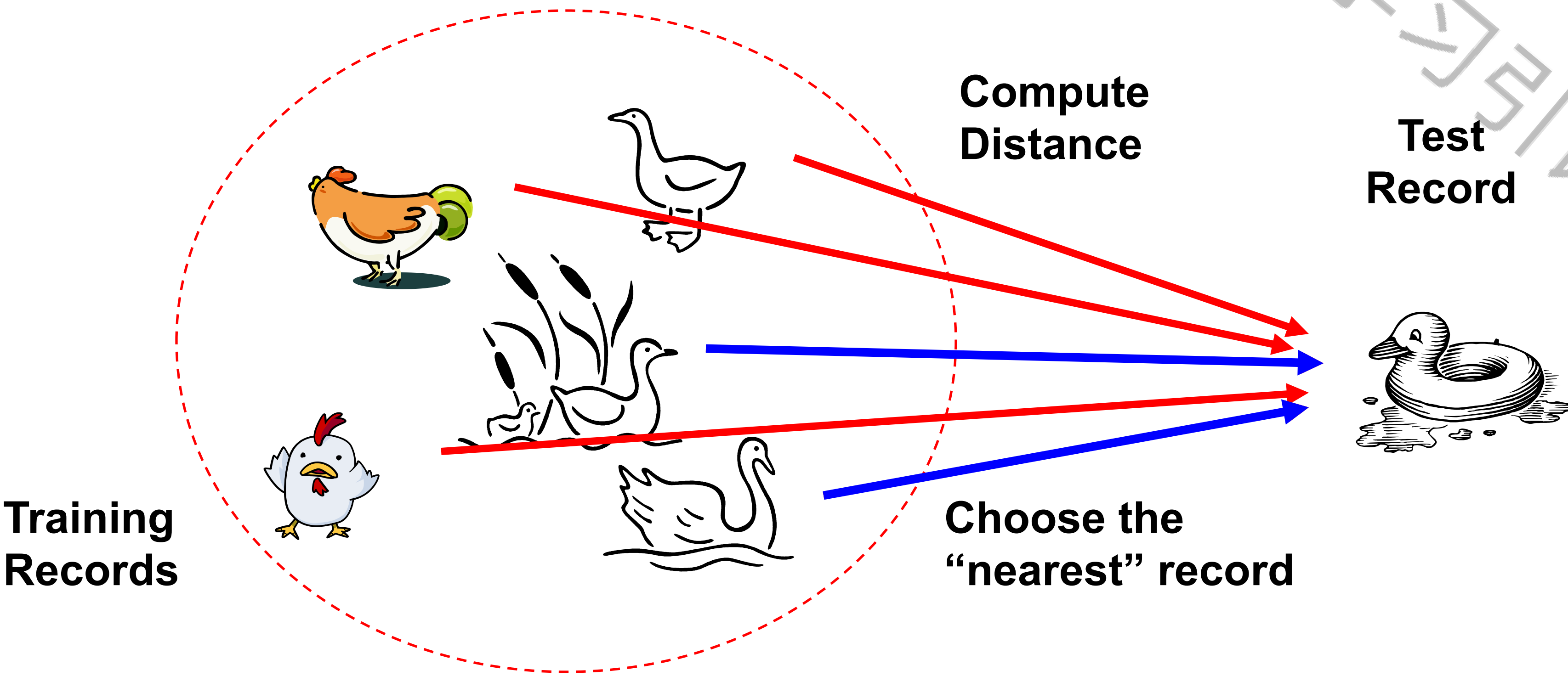
二、Classification Like Human

How to recognize a new object our brain?



二、Classification Like Human

How to recognize a new object our brain?

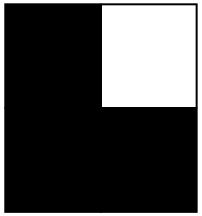


If it walks like a duck, quacks like a duck, then it’s probably a duck！

二、Classification Like Human

Step 1: represent the testing data point (x) in the vector space whose elements denote the ``features" .

属性	样本1	样本2
叫声	1	2
毛发	1	2
walk	0	1
…		

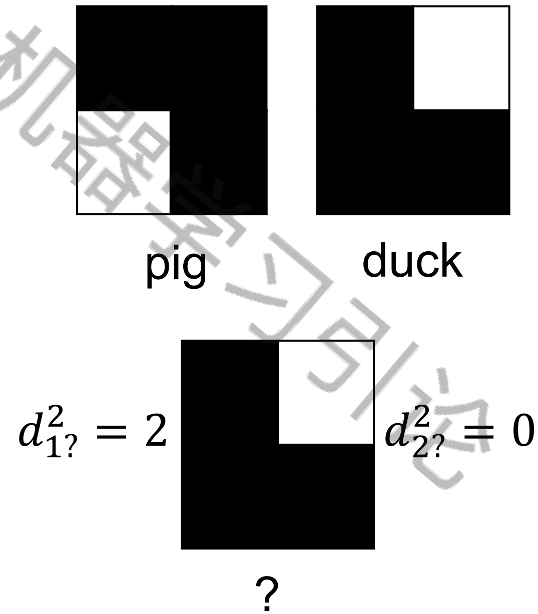


?

二、Classification Like Human

Step 1: represent the testing data point (x) in the vector space whose elements denote the ``features” .

Step 2: compute the distance between the testing data point and training data points.

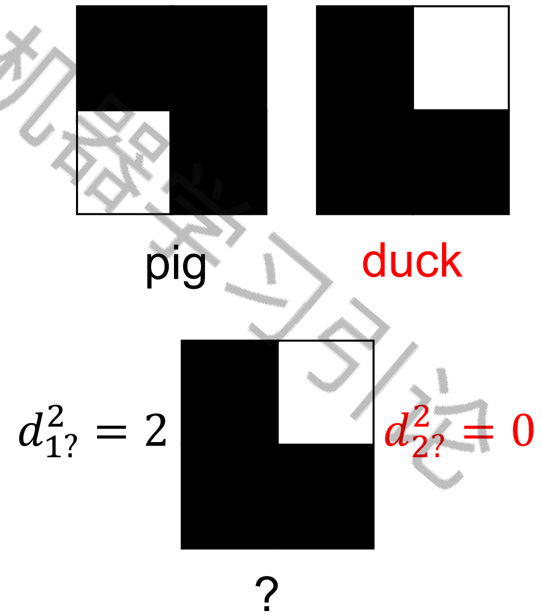


二、Classification Like Human

Step 1: represent the testing data point (x) in the vector space whose elements denote the ``features” .

Step 2: compute the distance between the testing data point and training data points.

Step 3: assign the sample to the nearest subject.

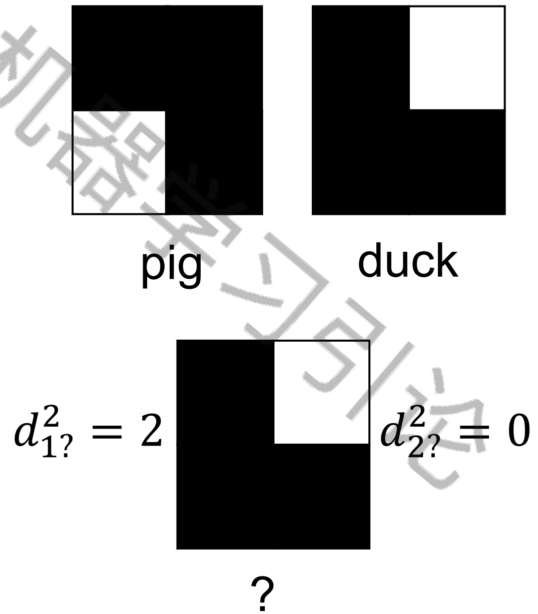


二、Classification Like Human

Step 1: represent the testing data point (x) in the vector space whose elements denote the ``features" .

Step 2: compute the distance between the testing data point and training data points.

Step 3: assign the sample to the nearest subject.



The nearest neighbor classifier (1NN)

Problems？ Limitation？

二、Classification Like Human

Step 1: represent the testing data point (x) in the vector space whose elements denote the ``features'' .

Step 2: compute the distance between the testing data point and training data points.

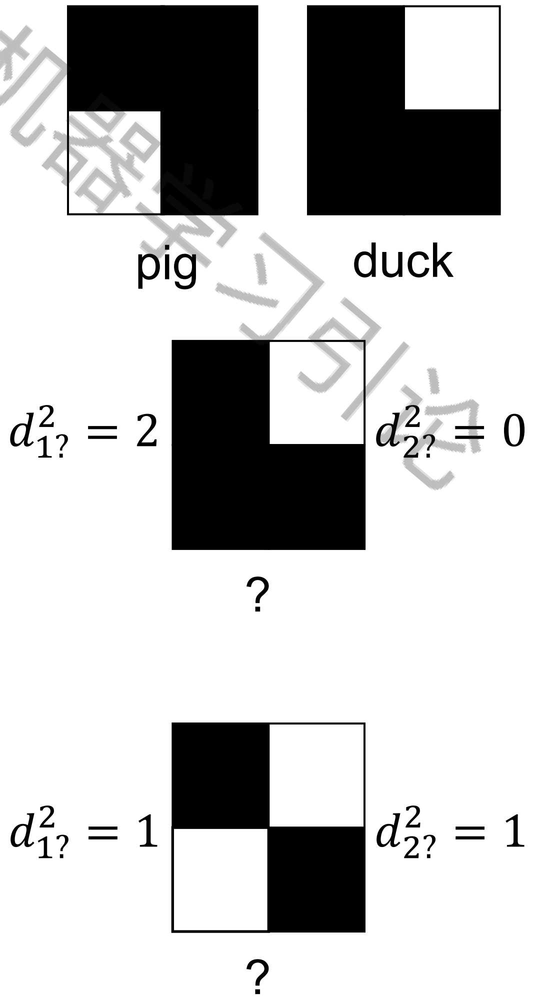
Step 3: assign the sample to the nearest subject.

The nearest neighbor classifier (1NN)

Problems:

The training data are sufficiently distinct with each other. Insufficient robustness to noises.

…

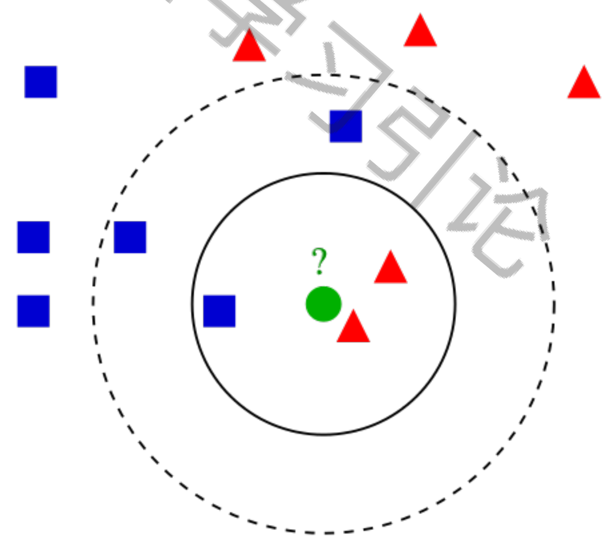


二、Classification Like Human

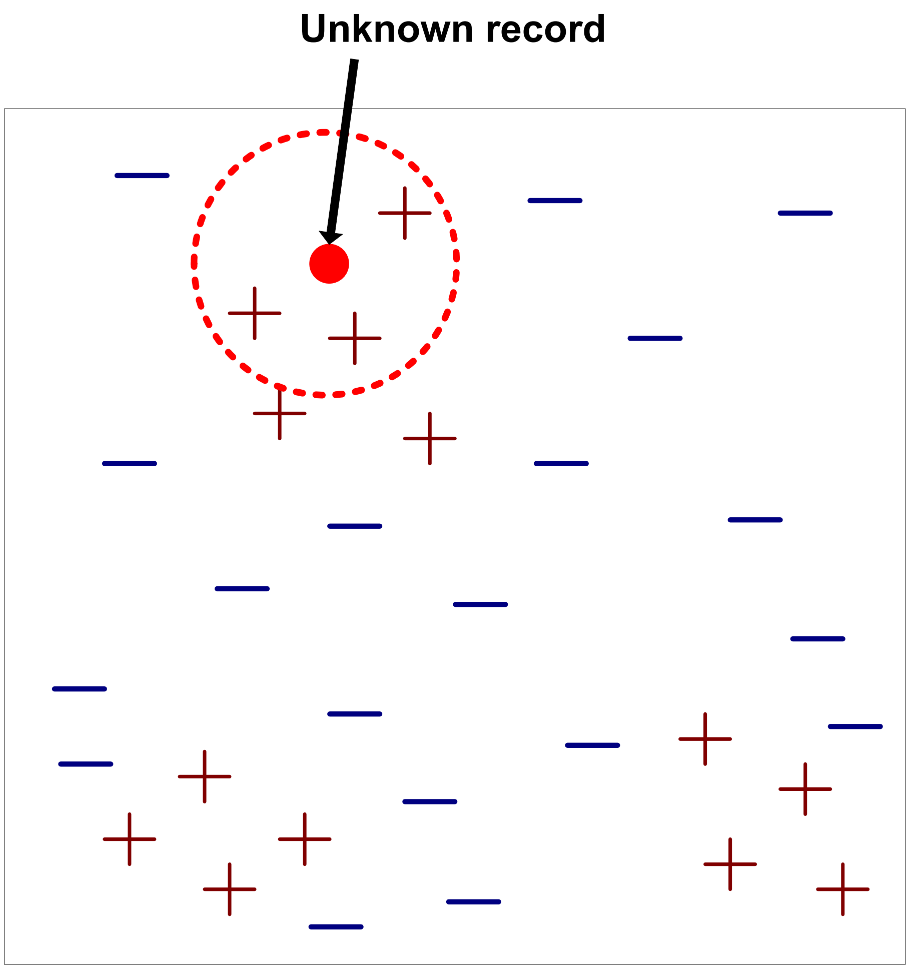
Prob1: The training data are sufficiently distinct with each other. Insufficient robustness to noises.

Sol: using k-nearest neighbor + max voting.

k-nearest neighbor classifier !



二、Classification Like Human



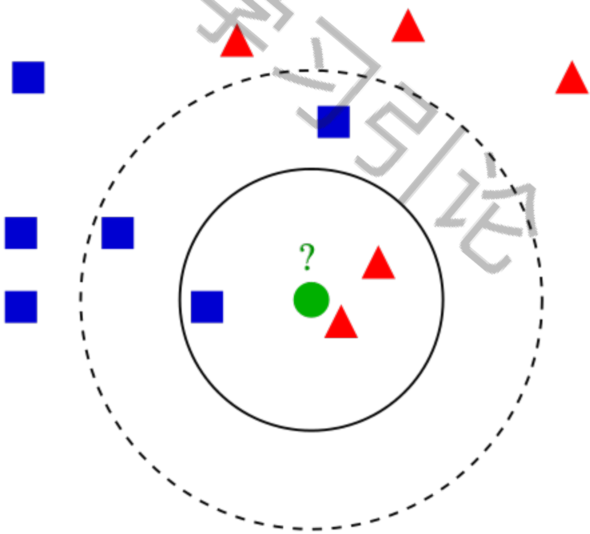
- Requires three things
  - The set of stored patterns
  - Distance Metric to compute distance between patterns
  - The value of *k*, the number of nearest neighbors to retrieve

- To classify an unknown pattern:
  - Compute distance to other training patterns
  - Identify *k* nearest neighbors
  - Use class labels of nearest neighbors to determine the class label of unknown pattern (e.g., by taking majority vote)

二、Classification Like Human

The KNN classifier:

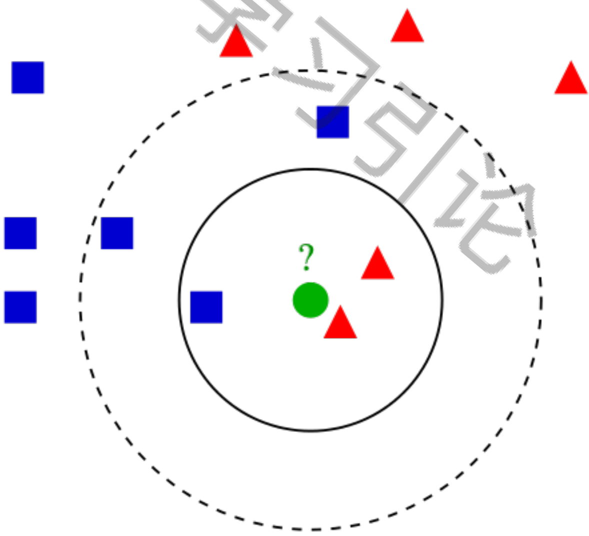
- Prob: does not take the distance into the consideration of voting.
- Sol: ?



二、Classification Like Human

The KNN classifier:

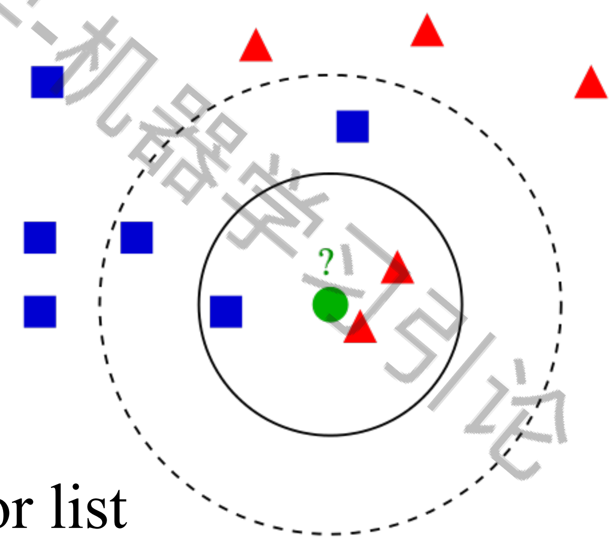
- Prob: does not take the distance into the consideration of voting.
- Sol: Weighting by the distance!



二、Classification Like Human

- Compute distance between two points:
  - Euclidean distance

$$d(p,q)=\sqrt{\sum_i (p_i - q_i)^2}$$



- Determine the class from the nearest neighbor list
  - take the majority vote of class labels among the k-nearest neighbors. $\begin{bmatrix}2\\1\end{bmatrix}$
  - Weight the vote according to distance
    - weight factor, w = $1/d^2$
    - $\begin{bmatrix}2\times0.8\\1\times0.1\end{bmatrix}=\begin{bmatrix}1.6\\0.1\end{bmatrix}$

同时考虑领域内类内样本数量和距离！

提纲

一．Review

二．Classification Like Human – KNN classifier

三．Performance Metric

四．Model Selection and Significant Test

五．Normalization

三、Performance Metric

怎么评价分类器的性能好坏？

三、Performance Metric

怎么评价分类器的性能好坏？

Accuracy or misclassification error (most popular)

- Error = classifying a record as belonging to one class when it belongs to another class.
- Error rate = percent of misclassified records out of the total records in the validation data

$$g = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, p = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}$$

$$accuracy = \frac{|g == p|}{|g|} = 33.33\%$$

三、Performance Metric

怎么评价分类器的性能好坏？

Confusion matrix（混淆矩阵），考虑二分类问题：

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	201	85
0	25	2689

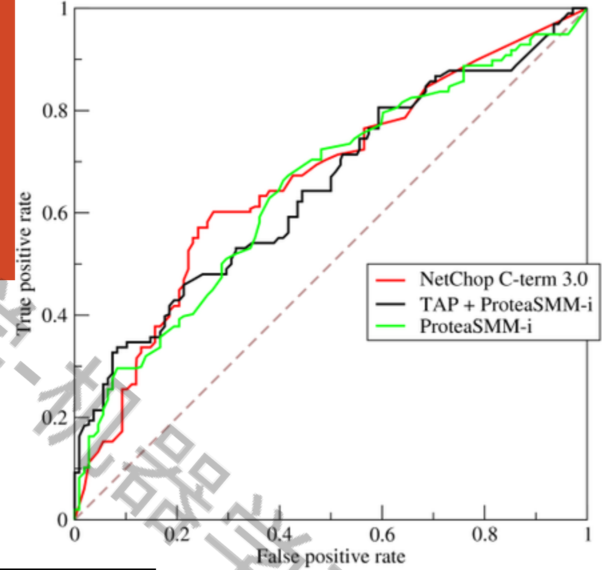
201 1’ s correctly classified as “1”

85 1’ s incorrectly classified as “0”

25 0’ s incorrectly classified as “1”

2689 0’ s correctly classified as “0”

三、Performance Metric



怎么评价分类器的性能好坏？

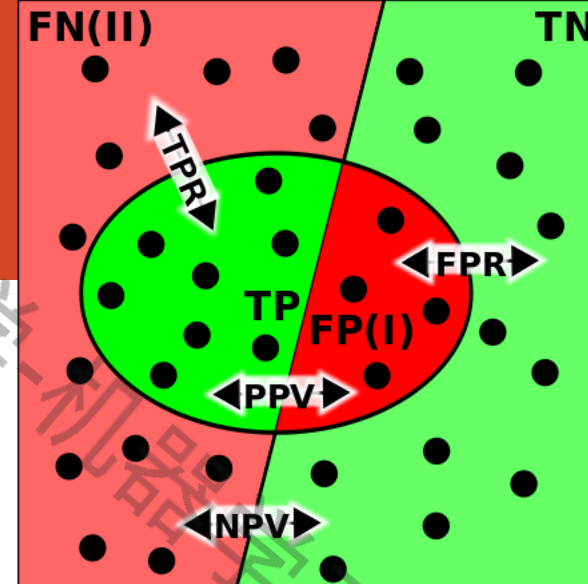
Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	201	85
0	25	2689

Overall error rate = (25+85)/3000 = 3.67%

Accuracy = 1 – err = (201+2689) = 96.33%

How about multiple-class problem?

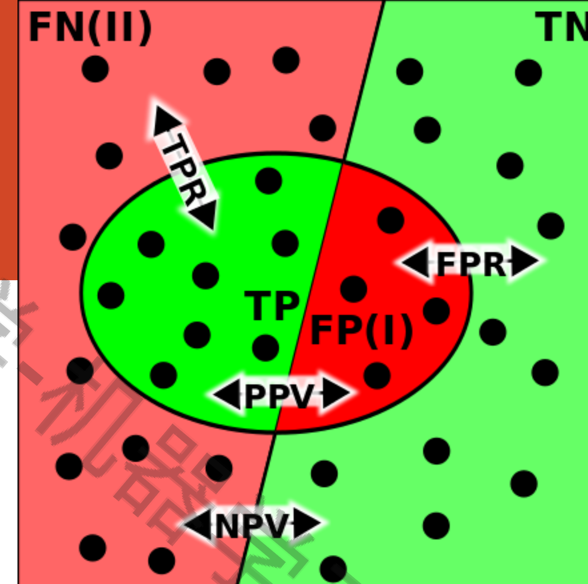
三、Performance Metric



怎么评价分类器的性能好坏？

		True condition			
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$ F_1 score $= \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	True negative rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$	

三、Performance Metric



怎么评价分类器的性能好坏？

		True condition				
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score $= \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	True negative rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		

Q：什么情况下，假阳性/假阴性比更重要？

提纲

一. Review

二. Classification Like Human – KNN classifier

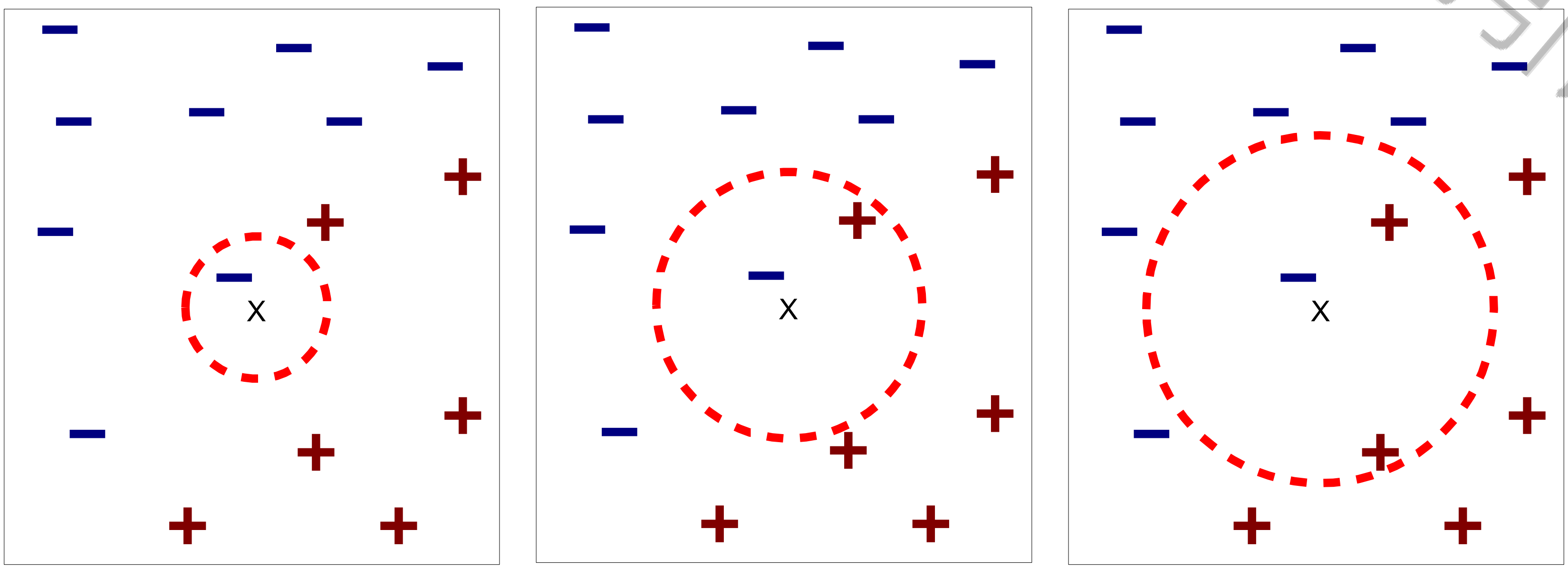
三. Performance Metric

四. Model Selection and Significant Test

五. Normalization

四、Model Selection and Significant Test

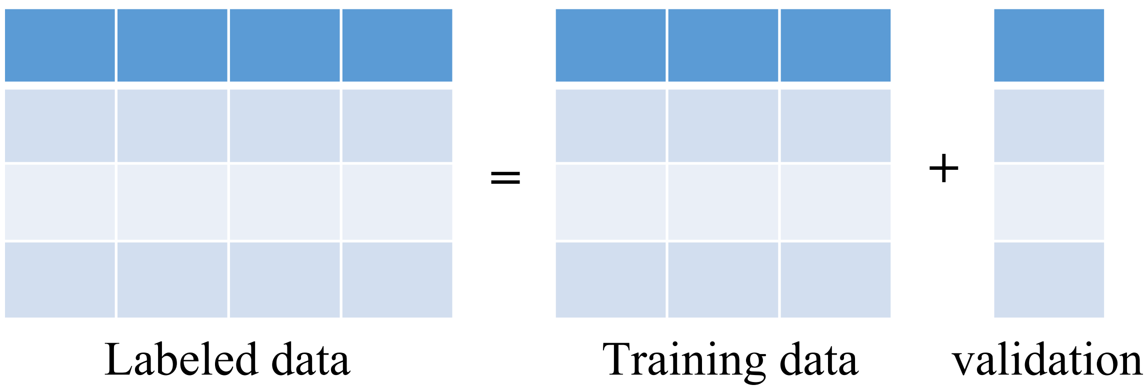
- Prob: choosing the value of k:
  - If k is too small, sensitive to noise points
  - If k is too large, neighborhood may include points from other classes



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

四、Model Selection and Significant Test

Prob: choosing the value of k, i.e., model selection.
Sol: tuning parameters using validation subset.



四、Model Selection and Significant Test

How to prove the method is good in statistics.

- Holdout method
  - Partition: Training-and-testing
    - Given data is randomly partitioned into two independent sets
    - Training set (e.g., 2/3) for model construction
    - Test set (e.g., 1/3) for accuracy estimation
    - Unbiased, efficient. But require a large number of samples
    - used for data set with large number of samples
  - Random sampling: a variation of holdout
    - Repeat holdout k times, accuracy = avg. of the accuracies obtained

四、Model Selection and Significant Test

How to prove the method is good in statistics.

- Cross-validation (*k*-fold, where k = 10 is most popular)
  - Randomly partition the data into *k mutually exclusive* subsets, each approximately equal size
  - At *i*-th iteration, use D_i as test set and others as training set
  - Leave-one-out: k folds where k = # of tuples, for small sized data
  - Stratified cross-validation: folds are stratified so that class dist. in each fold is approx. the same as that in the initial data

提纲

一．Review

二．Classification Like Human – KNN classifier

三．Performance Metric

四．Model Selection and Significant Test

五．Normalization

五、Normalization

- Prob: Scaling issues
  - Attributes may have to be scaled to prevent distance measures from being dominated by one of the attributes
  - Example:
    - height of a person may vary from 1.5m to 1.8m
    - weight of a person may vary from 90lb to 300lb
    - income of a person may vary from $10K to $1M

五、Normalization

- Prob: Scaling issues
  - Attributes may have to be scaled to prevent distance measures from being dominated by one of the attributes
  - Example:
    - height of a person may vary from 1.5m to 1.8m
    - weight of a person may vary from 90lb to 300lb
    - income of a person may vary from \$10K to \$1M
- Sol: Normalization

五、Normalization

Name	Formula	Use
Standard score	$\frac{X-\mu}{\sigma}$	Normalizing errors when population parameters are known. Works well for populations that are normally distributed
Student's t-statistic	$\frac{X-\overline{X}}{s}$	Normalizing residuals when population parameters are unknown (estimated).
Studentized residual	$\frac{\hat{\epsilon}_i}{\hat{\sigma}_i}=\frac{X_i-\hat{\mu}_i}{\hat{\sigma}_i}$	Normalizing residuals when parameters are estimated, particularly across different data points in regression analysis.
Standardized moment	$\frac{\mu_k}{\sigma^k}$	Normalizing moments, using the standard deviation σ as a measure of scale.
Coefficient of variation	$\frac{\sigma}{\mu}$	Normalizing dispersion, using the mean μ as a measure of scale, particularly for positive distribution such as the exponential distribution and Poisson distribution.
Feature scaling	$X'=\frac{X-X_{\min}}{X_{\max}-X_{\min}}$	Feature scaling is used to bring all values into the range [0,1]. This is also called unity-based normalization. This can be generalized to restrict the range of values in the dataset between any arbitrary points a and b using $X'=a+\frac{(X-X_{\min})(b-a)}{X_{\max}-X_{\min}}$.

Take home

The NN classifier:

- Prob: The training data are sufficiently distinct with each other. Insufficient robustness to noises.
- Sol: using k-nearest neighbor + max voting.

The KNN classifier:

- Prob: does not take the distance into the consideration of voting.
- Sol: Weighting by the distance!
- Prob: Choosing the value of k, i.e., model selection
- Sol: split the labeled data into training set and validation set.
- Prob: How to prove the method is good in statistics.
- Sol: Holdout method/Cross-validation
- Prob: Scaling issues
- Sol: Normalization

Test Questions

- Q1: What is the classification? How to perform classification by human? And what is the simplest way?
- Q2: What problem of 1NN is addressed by kNN?
- Q3: How to (why) incorporate the distance into classical kNN? And what will be benefited from it?
- Q4: How to solve the scaling issue faced by KNN?
- Q5: How to evaluate the performance of a classifier?
- Q6: What is model selection? How to solve this issue?

Any other demerits?

Next Course

Kernel and SVM

Q&A

THANKS!