Molecular Biomedical Informatics

分 子 生 醫 資 訊 實 驗 室

Machine Learning and Bioinformatics

機 器 學 習 與 生 物 資 訊 學

Feature

Various problems/techniques

Feature types

Nominal, ordinal, interval and ratio

Nominal feature

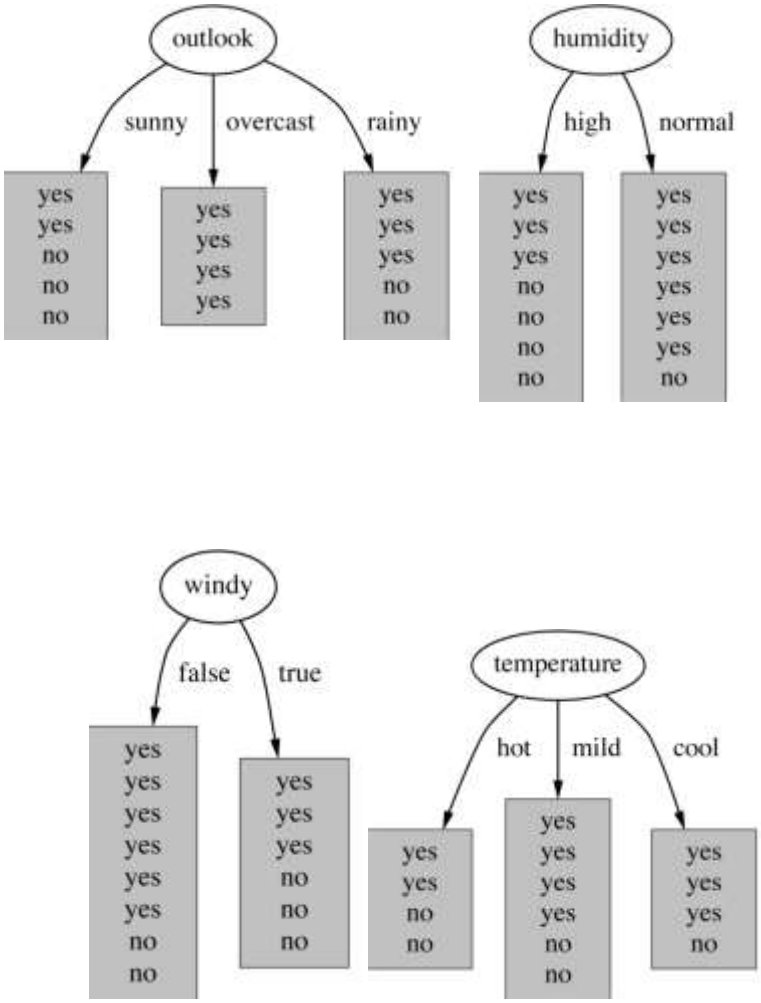
- Values are distinct symbols
  - values themselves serve only as labels or names
  - nominal comes from the Latin word for name
- Example
  - feature “category” of a stock
  - such as “Food and Beverage / 食品工業”, “Chemicals/化學工業”, “Automotive/汽車工業”, “Transportation/航運業”, “Electronics/電子工業”, …
- No relation is implied among nominal values (no ordering or distance measure)
- Only equality tests can be performed

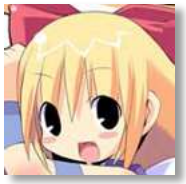
Decision tree

- Strategy: top down
- Recursive divide-and-conquer fashion
  - select a feature for root node
  - create a branch for each possible feature value
  - split instances into subsets
  - repeat above steps recursively for each branch
- Stop if all instances have the same class

Which feature?

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No





How to

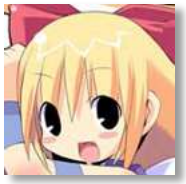
use nominal features in KNN

Ordinal features

- Impose order on values
  - but no distance between values defined
  - such as “temperature” in the weather data, where “hot” > “mild” > “cool”
- Addition and subtraction don’t make sense
- Distinction between nominal and ordinal not always clear (e.g. “outlook” in the weather data)

Interval and ratio features

- Interval features are not only ordered but measured in fixed and equal units
  - difference of two values makes sense; sum or product doesn't, ex: "birth year"
  - "temperature" expressed in degrees Fahrenheit
- Ratio features are treated as real numbers
  - all mathematical operations are allowed, ex: "age"
- Both can be called numeric features

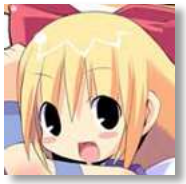


What's

the difference between interval and ratio features

Any Questions?

about numeric features



How to

use numeric features in decision tree

Dealing with numeric features

- Discretize numeric values
  - sort instances according to the feature's values
  - breaks where class changes (majority class)
  - this minimizes the error

64	65	68	69	70	71	72	72	75	75	80	81	83	85
Yes	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	No

The problem of overfitting

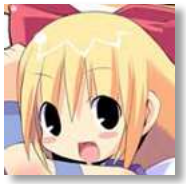
- The above procedure is very sensitive to noise
  - one instance with an incorrect class label probably produces a separate interval
- A time stamp results in an error-free, but highly branched, feature
- Enforce minimum number of instances of the majority class per interval (ex: 3)

64	65	68	69	70	71	72	72	75	75	80	81	83	85
Yes	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	No

Missing value

Missing value

- Frequently indicated by out-of-range entries
  - unknown, unrecorded, irrelevant, …
- Causes
  - malfunctioning equipment
  - changes in experimental design
  - collation of different datasets
  - measurement not possible
- Missing value may have significance in itself
  - missing test in a medical examination
- Trivial in decision tree



But

how to solve it in KNN

Inaccurate values

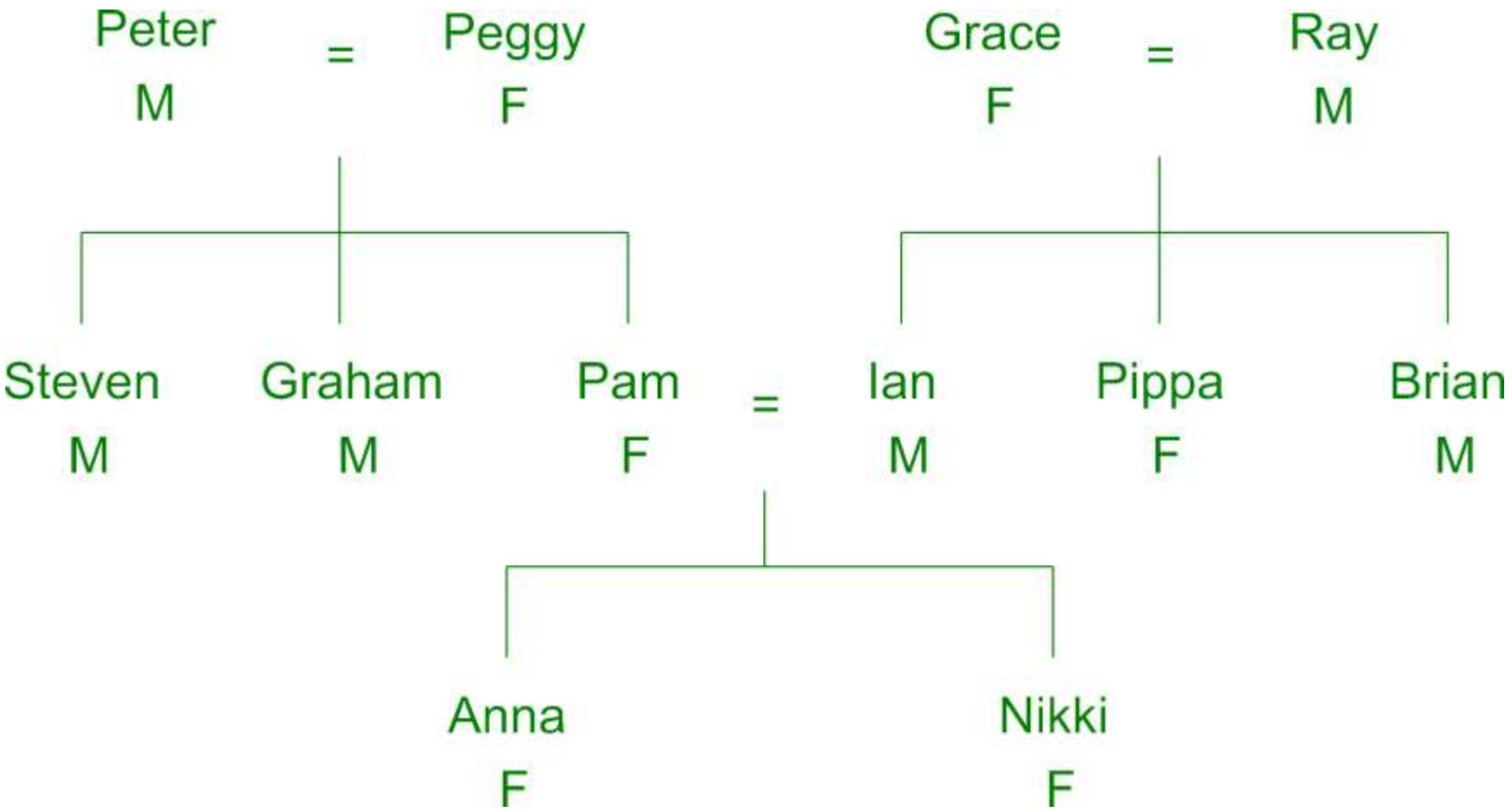
- Stale data, ex: “age”
- Typographical errors
- Errors may be deliberate, ex: “phone”

- A research problem of machine learning
  - noise/outlier detection

Paired instances

Watch out the instance definition when extracting features

A family tree



The family tree represented as a table

Name	Gender	Parent1	Parent2
Peter	Male	?	?
Peggy	Female	?	?
Steven	Male	Peter	Peggy
Graham	Male	Peter	Peggy
Pam	Female	Peter	Peggy
Ian	Male	Grace	Ray
Pippa	Female	Grace	Ray
Brian	Male	Grace	Ray
Anna	Female	Pam	Ian
Nikki	Female	Pam	Ian

The “sister-of” relation

First person				Second person				Relation
Name	Gender	Parent1	Parent2	Name	Gender	Parent1	Parent2	
Steven	Male	Peter	Peggy	Pam	Female	Peter	Peggy	Yes
Graham	Male	Peter	Peggy	Pam	Female	Peter	Peggy	Yes
Ian	Male	Grace	Ray	Pippa	Female	Grace	Ray	Yes
Brian	Male	Grace	Ray	Pippa	Female	Grace	Ray	Yes
Anna	Female	Pam	Ian	Nikki	Female	Pam	Ian	Yes
Nikki	Female	Pam	Ian	Anna	Female	Pam	Ian	Yes
All the rest								No

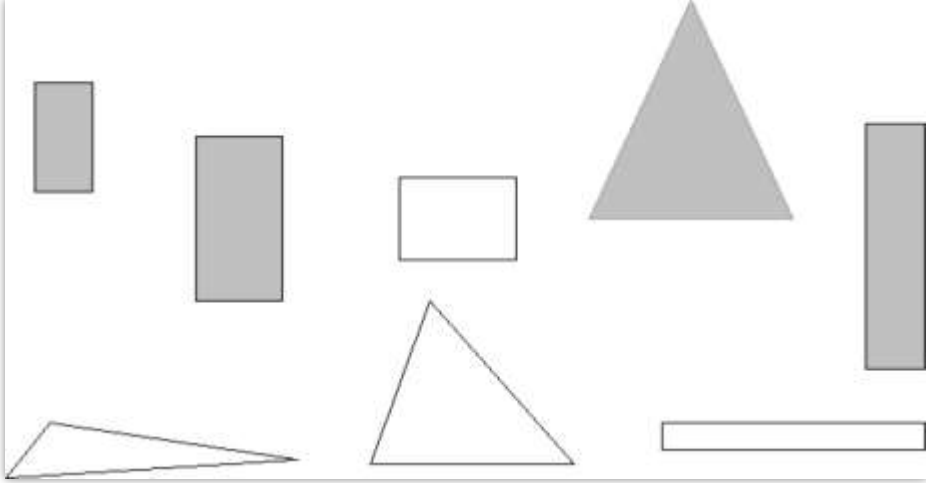
The "sister-of" relation

- If second person's gender = female and the first person's parent = the second person's parent
- The order of individuals
  - change features to order-independent
  - generate new features to fit the requirement ← domain knowledge is required
  - generate reverse training samples ← instance-level solution
  - change the distance function ← kernel-level solution

$$\sum class_i \times d(query, train_i)\text{, where } d()$$

$$= \min(|query - train_i|, |reverse(query) - train_i|)$$

Further problems



- "Ancestor-of"
- "Standing-vs.-lying"
  - if width > height then lying; if height > width then standing ← no machine learning tools can do that
  - generating new features is required
- How do you know if stock prediction requires such derived features?

Class

Also influence the learning behaviors

Summary

- Intrinsic characteristics of features
- Problem/instance definition
  - feature arrangement
  - class assignment
- In stock
  - change/distance quantities
  - normalized quantities (to an instance or to all instances)
  - 股票技術分析

Today’s exercise

Feature encoding

Generate features from your raw data. Send TA Lin a report before 23:59 10/23 (Wed).

Appendix

Feature types used in practice

- Most schemes accommodate just two levels of measurement: categorical and numerical
- Categorical features are also called “enumerated” or “discrete”, but “enumerated” and “discrete” imply order
- A special feature type is dichotomy (“boolean”)
- Numeric features are also called “continuous”, but “continuous” implies mathematical continuity

Predicting protein–protein interactions based only on sequences information

Juwen Shen[†], Jian Zhang[†], Xiaomin Luo[†], Weiliang Zhu[†‡], Kunqian Yu[†], Kaixian Chen[†], Yixue Li[§], and Hualiang Jiang[†‡¶]

[†]Center for Drug Discovery and Design, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, and Graduate School of Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai 201203, China; [‡]School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China; and [§]Bioinformation Center, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yue Yang Road, Shanghai 200031, China

Edited by Michael Levitt, Stanford University School of Medicine, Stanford, CA, and approved December 28, 2006 (received for review September 8, 2006)

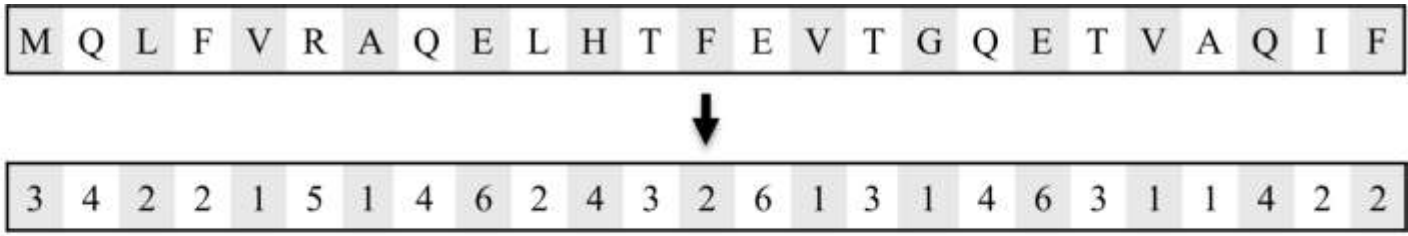
Protein–protein interactions (PPIs) are central to most biological processes. Although efforts have been devoted to the development of methodology for predicting PPIs and protein interaction networks, the application of most existing methods is limited because they need information about protein homology or the interaction marks of the protein partners. In the present work, we propose a method for PPI prediction using only the information of protein sequences. This method was developed based on a learning algorithm-support vector machine combined with a kernel function and a conjoint triad feature for describing amino acids. More than 16,000 diverse PPI pairs were used to construct the universal model. The prediction ability of our approach is better than that of other sequence-based PPI prediction methods because it is able to predict PPI networks. Different types of PPI networks have been effectively mapped with our method, suggesting that, even with only sequence information, this method could be applied to the exploration of networks for any newly discovered protein with unknown biological relativity. In addition, such supplementary experimental information can enhance the prediction ability of the method.

conjoint triad | support vector machine

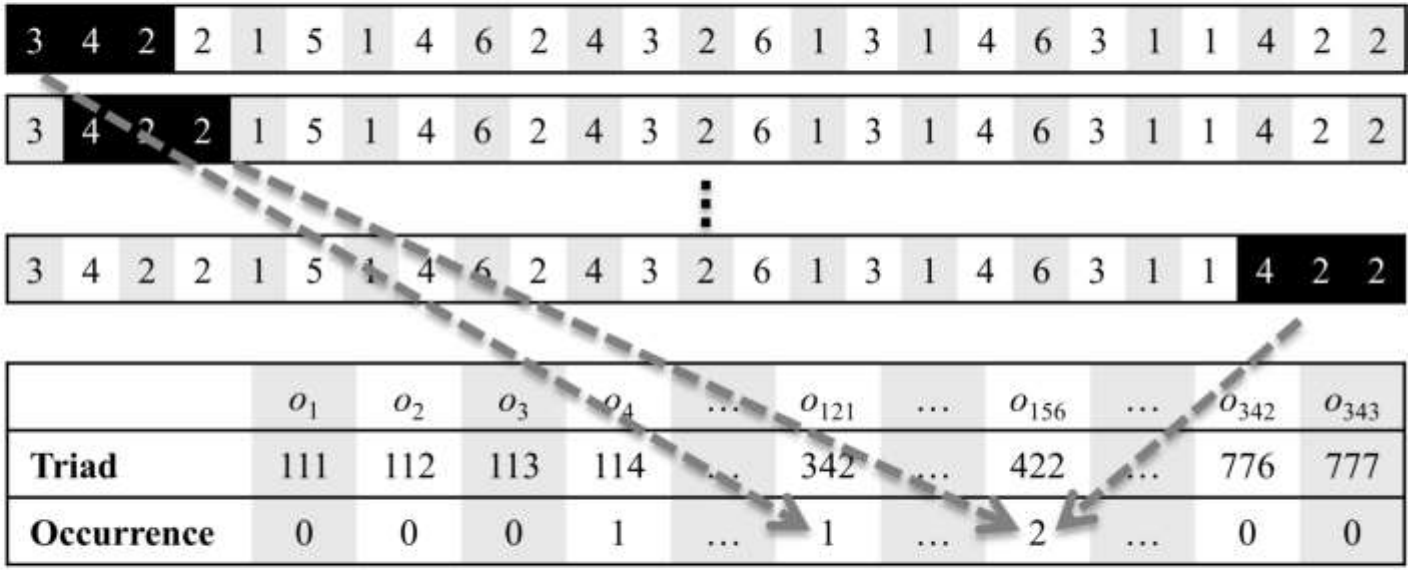
The molecular bases of cellular operations are sustained largely by different types of interactions among proteins. Thus, a major goal of functional genomics is to determine protein interaction networks for whole organisms (1). However, approaches (13). Several binding reaction-detected methods, based on the presumption that the binding of one protein to another provokes a variety of biophysical changes, have been developed (14). These technologies recently identified hundreds of potentially interacting proteins and complexes in several species such as yeast (15), *Drosophila* (16), and *Helicobacter pylori* (17). Ulrich *et al.* (18) presented a large-scale two-hybrid map of >3,000 putative human PPIs. These data will serve as an important source of information regarding individual protein partners and offer preliminary insight into the global molecular organization of human cells.

However, current PPI pairs obtained with experimental methods cover only a fraction of the complete PPI networks (19). Therefore, computational methods for the prediction of PPIs have an important role (20). A number of computational methods have been developed for the prediction of PPIs. Computational methods based on genomic information, such as phylogenetic profiles, predict PPIs by accounting for the pattern of the presence or absence of a given gene in a set of genomes (21, 22). The main limitation of these approaches is that they can be applied only to completely sequenced genomes, which is the precondition to rule out the absence of a given gene. Similarly, they cannot be used with the essential proteins that are common to most organisms (23). The prediction of functional relationships between two proteins according to their corresponding adjacency of genes is another popular approach. This method is directly applicable only to bacteria, in which the genome order

Step 1: Transform the amino acid sequence into a sequence of amino acid groups.



Step 2: Scan all triads along the sequence of amino acid groups, and count the triads in the occurrence vector **O**.



	o_1	o_2	o_3	o_4	…	o_{121}	…	o_{156}	…	o_{342}	o_{343}
Triad	111	112	113	114	…	342	…	422	…	776	777
Occurrence	0	0	0	1	…	1	…	2	…	0	0

Occurrence vector $\mathbf{O}=<o_1, o_2, \ldots, o_{343}>$

Step 3: Convert occurrence vector into significance vector.

Significance vector $\mathbf{S}=<s_1, s_2, \ldots, s_{343}>$, where s_i is obtained by estimating the probability of observing less occurrences of the i-th triad than the one that is actually observed (o_i) in a background distribution.

Yu *et al. BMC Bioinformatics* 2010, **11**:167
http://www.biomedcentral.com/1471-2105/11/167

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Predicting protein-protein interactions in unbalanced data using the primary structure of proteins

Chi-Yuan Yu[1], Lih-Ching Chou[1] and Darby Tien-Hao Chang*[2]

Abstract

Background: Elucidating protein-protein interactions (PPIs) is essential to constructing protein interaction networks and facilitating our understanding of the general principles of biological systems. Previous studies have revealed that interacting protein pairs can be predicted by their primary structure. Most of these approaches have achieved satisfactory performance on datasets comprising equal number of interacting and non-interacting protein pairs. However, this ratio is highly unbalanced in nature, and these techniques have not been comprehensively evaluated with respect to the effect of the large number of non-interacting pairs in realistic datasets. Moreover, since highly unbalanced distributions usually lead to large datasets, more efficient predictors are desired when handling such challenging tasks.

Results: This study presents a method for PPI prediction based only on sequence information, which contributes in three aspects. First, we propose a probability-based mechanism for transforming protein sequences into feature vectors. Second, the proposed predictor is designed with an efficient classification algorithm, where the efficiency is essential for handling highly unbalanced datasets. Third, the proposed PPI predictor is assessed with several unbalanced datasets with different positive-to-negative ratios (from 1:1 to 1:15). This analysis provides solid evidence that the degree of dataset imbalance is important to PPI predictors.

Non-linear normalization

- An example of using 2 residue groups and bigrams
  - AAAAAATT
    - AA 5
    - AT 1
    - TA 0
    - TT 1
- The one occurrence of TT should be more significant than the one occurrence of AT
- The original encoding
  - 1:5 2:1 3:0 4:1

Triad significance

- o_i represents the number of the i-th type of triad observed, which is highly correlated to the distribution of amino acids
- The significance is defined by answering the following question
  - How rare is the number of observed occurrences considering the amino acid composition of the protein?
- We define s_i as the probability of observing less occurrences of the i-th triad than the one that is actually observed (o_i), which equals to 1 minus the p-value
- Estimate the importance of occurrence

Final encoding

- $s_i = \Pr(X_i < o_i)$, where X_i is a random variable representing the number of observations of the i-th triad in a background distribution of protein sequences
- A common practice to estimate X_i is to permute the original protein sequence many times while preserving its amino acid composition
- The previous example
  - `1:5  2:1  3:0  4:1`
  - `1:Pr(AA<5)  2:Pr(AT<1)  3:Pr(TA<0)  4:Pr(TT<1)`
  - `1:0.895  2:0.037  3:0.000  4:0.735`