Computational Intelligence: Methods and Applications

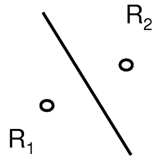
Lecture 33
Decision Tables & Information Theory

Włodzisław Duch
SCE, NTU, Singapore
Google: Duch

Knowledge from kNN?

Since all training vectors are used decision borders are complex and hard to understand in logical (symbolic terms).

Reduce the number of reference vectors R_i
one vector per class with Euclidean distance measure
<=> LDA linear machine solution!



Select prototype vectors by:
removing and checking the accuracy using the leave-one-out test
adding cluster center and adapting its positions to maximize L1O

Example: Wisconsin breast cancer dataset, simplest known description.
699 samples, 458 benign (65.5%) and 241 (34.5%) malignant cancer cases, 9 features - integers ranging 1-10.

IF $D(X,R_{681}) < 51.97$ THEN malignant, ELSE benign
Accuracy 97.4%, Specificity 98.8%, Sensitivity 96.8%, so far the best!

Decision tables

kNN is a local majority classifier.

Decision Tables is a similar algorithm that works only for nominal features: given a training data Table(**R**,C), try to reduce the number of features and the number of samples without decreasing accuracy.

Decision rule is:
find all R= reduced X, predict majority of R class.

Reducing features may make some vectors X identical.

Example: if only 1 feature is left, X_1=0 or 1, a number of original training data vectors R for each value of X_1 will have ω_1 and ω_2 labels.
If ω_1 is in majority for X_1=0, and ω_2 for X_1=1, then any new sample with X_1=0 (or X_1=1) is classified to ω_1 (or ω_2).

Search techniques are used to find reduced features: find and remove feature that has no influence (or that increases) accuracy, repeat.

Decision tables in WEKA/Yale

If reduced X does not match any R in the table, majority class is predicted.

Decision table algorithm looks for reduced set of features, this is similar to what the “rough set” programs do.

WEKA does selection of features and instances using information-theoretic measures, and search using best-first technique.

-S num Number of fully expanded non improving subsets to consider before terminating a best first search. (Default = 5)

-X num: Use cross validation to evaluate feature usefulness. Use number of folds = 1 for leave one out CV. (Default = leave one out CV)

-I: Use nearest neighbor instead of global table majority.

Concept of information

Information may be measured by the average amount of surprise of observing X (data, signal, object).

1. If *P*(X)=1 there is no surprise, so *s*(X)=0
2. If *P*(X)=0 then this is a big surprise, so *s*(X)=∞.
3. If two observations X, Y are independent than *P*(X,Y)=*P*(X)*P*(Y),

   but the amount of surprise should be a sum *s*(X,Y)=*s*(X)+*s*(Y).

The only suitable surprise function that fulfills these requirements is ...

$$s\left(X\right)=\lg\frac{1}{P\left(X\right)}=-\lg P\left(X\right)$$

The average amount of surprise is called information or entropy.
Entropy is a measure of disorder, information is the change in disorder.

Information

Information derived from observations of variable *X* (vector variable, signal or some object) that has *n* possible values is thus defined as:

$$H(X)=E\left[\lg\frac{1}{P\left(X\right)}\right]=-\sum_{i=1}^{n}P\left(X^{(i)}\right)\lg P\left(X^{(i)}\right)\geq 0$$

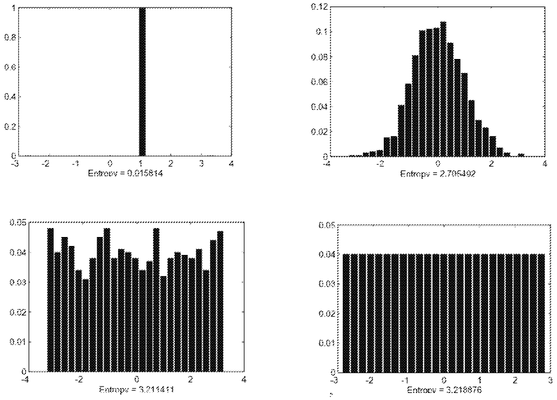
If the variable *X* is continuous with distribution *P*(x) an integral is taken instead of the sum: $H(X)=-\int P\left(x\right)\lg P\left(x\right)dx$

What type of logarithm should be used? Consider binary event with *P*(X$^{(1)}$)=*P*(X$^{(2)}$)=0.5, like tossing a coin: how much do we learn each time? A bit. Exactly one bit. Taking lg$_2$ will give:

$$H(X)=-0.5\lg_2 0.5-0.5\lg_2 0.5=1$$

Distributions

For a scalar variable *P*(X$^{(i)}$) may be displayed in form of a histogram, and information (entropy) calculated for each histogram.



Joint information

Other ways of introducing the concept of information start from the number of bits needed to code a signal.

Suppose now that two variables, X and Y, are observed.
Joint information is:

$$H(X,Y)=E\left[\lg\frac{1}{P\left(X,Y\right)}\right]=-\sum_{i,j=1}^{n,m}P\left(X^{(i)},Y^{(j)}\right)\lg P\left(X^{(i)},Y^{(j)}\right)$$

For two uncorrelated features this is equal to:

$$H(X,Y)=-\sum_{i,j=1}^{n,m}P\left(X^{(i)}\right)P\left(Y^{(j)}\right)\left(\lg P\left(X^{(i)}\right)+\lg P\left(Y^{(j)}\right)\right)$$

$$=H\left(X\right)+H\left(Y\right)$$

Since: $\sum_{i=1}^{n}P\left(X^{(i)}\right)=\sum_{i=1}^{m}P\left(Y^{(i)}\right)=1$

Conditional information

If the value of *Y* variable is fixed and *X* is not quite independent conditional information (average “conditional surprise”) may be useful:

$$H(X\,|\,Y)=E_{XY}\left[\lg\frac{1}{P(X\,|\,Y)}\right]$$

$$=-\sum_{i,j=1}^{n,m}P\left(X^{(i)},Y^{(j)}\right)\lg P\left(X^{(i)}\,|\,Y^{(j)}\right)$$

$$=-\sum_{j=1}^{m}P\left(Y^{(j)}\right)\sum_{i=1}^{n}P\left(X^{(i)}\,|\,Y^{(j)}\right)\lg P\left(X^{(i)}\,|\,Y^{(j)}\right)$$

$$=\sum_{j=1}^{m}P\left(Y^{(j)}\right)H\left(X\,|\,Y^{(j)}\right)$$

Prove that $H(X\,|\,Y)=H(X,Y)-H(Y)$

Mutual information

The information in the X variable that is shared with Y is defined as:

$$MI(X;Y)=H(X)-H(X\,|\,Y)=$$

$$=-\sum_{i=1}^{n}P\left(X^{(i)}\right)\lg P\left(X^{(i)}\right)+\sum_{i,j=1}^{n,m}P\left(X^{(i)},Y^{(j)}\right)\lg P\left(X^{(i)}\,|\,Y^{(j)}\right)$$

$$=-\sum_{i,j=1}^{n,m}P\left(X^{(i)},Y^{(j)}\right)\lg P\left(X^{(i)}\right)+\sum_{i,j=1}^{n,m}P\left(X^{(i)},Y^{(j)}\right)\lg P\left(X^{(i)}\,|\,Y^{(j)}\right)$$

$$=\sum_{i,j=1}^{n,m}P\left(X^{(i)},Y^{(j)}\right)\lg\frac{P\left(X^{(i)}\,|\,Y^{(j)}\right)}{P\left(X^{(i)}\right)}\frac{P\left(Y^{(j)}\right)}{P\left(Y^{(j)}\right)}$$

$$=\sum_{i,j=1}^{n,m}P\left(X^{(i)},Y^{(j)}\right)\lg\frac{P\left(X^{(i)},Y^{(j)}\right)}{P\left(X^{(i)}\right)P\left(Y^{(j)}\right)}$$

Kullback-Leibler divergence

If two distributions for X variable are compared, their divergence is expressed by:

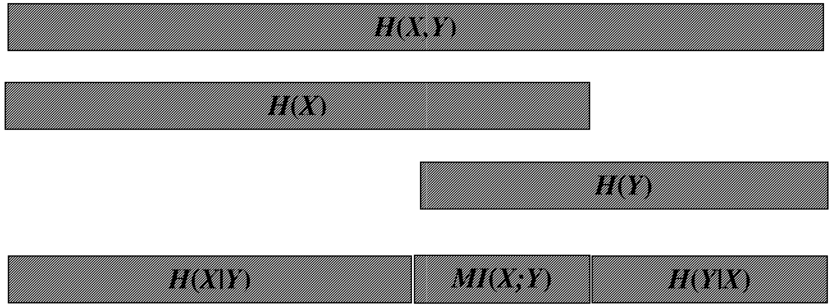
$$D_{KL}(p\,\|\,q)=\sum_{i=1}^{n}p\left(X^{(i)}\right)\lg\frac{p\left(X^{(i)}\right)}{q\left(X^{(i)}\right)}=E\left[\lg\frac{p\left(X^{(i)}\right)}{q\left(X^{(i)}\right)}\right]$$

KL divergence is the expected value of the ratio of two distributions, it is non-negative, but not symmetric in *p*, *q*, so it is not a distance, although sometimes it is referred to as “KL-distance”.

Mutual information is equal to the KL divergence between joint and product (independent) distributions:

$$MI(X;Y)=D_{KL}\left(P(X,Y)\,\|\,P(X)P(Y)\right)$$

Graphical relationships



H(*X*,*Y*)

H(*X*)

H(*Y*)

H(*X*|*Y*)

MI(*X*;*Y*)

H(*Y*|*X*)

So the total joint information is (prove it!)

$$H(X,Y)=MI(X;Y)+H(X\,|\,Y)+H(Y\,|\,X)$$

$$H(X,Y)=H(X)+H(Y)-MI(X;Y)$$