Feature Selection

1 General Approaches

- Finding a minimal subset of features that separate all vectors (class-independent).
- Searching the lattice of subsets of the set of features to find the subset that best represents the class distribution (computationaly intractable).
- Ranking: order features by their class discrimination power (for each term independently of the other terms, i.e. greedy search)
- Scheme-specific methods (e.g. attribute selection used in ID3)

2 Similarity-based attribute selection

2.1 Similarity (distance) measures

- *Euclidean distance*:
  $D(X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2}$
- *Cosine similarity* (dot product when normalized to unit length):
  $Sim(X,Y) = x_1.y_1 + x_2.y_2 + ... + x_n.y_n$
- Number of differences for nominal (boolean) attributes:
  $D(X,Y) = \sum_1^n d(x_i, y_i)$,
  where $d(x_i, y_i) = 0$ if $x_i = y_i$ and 1 otherwise.

2.2 Similarity-based attribute selection algorithm

- For each vector find the nearest neighbors (the closest vectors according to the distance measure) of the same and different classes – "near hits" and "near misses".
- If a near hit has a different value for a certain attribute then that attribute appears to be irrelevant and its weight should be decreased.
- For near misses, the attributes with different values are relevant and their weights should be increased.
- Algorithm: Start with equal weights for all attributes and do the weight adjustment, as explained above. This allows ordering attributes by relevance.

2.3 Example

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
X	x_1	x_2	x_3	x_4	y
X_1	sunny	hot	high	weak	no
X_2	sunny	hot	high	strong	no
X_3	overcast	hot	high	weak	yes
X_4	rain	mild	high	weak	yes
X_5	rain	cool	normal	weak	yes
X_6	rain	cool	normal	strong	no
X_7	overcast	cool	normal	strong	yes
X_8	sunny	mild	high	weak	no
X_9	sunny	cool	normal	weak	yes
X_{10}	rain	mild	normal	weak	yes
X_{11}	sunny	mild	normal	strong	yes
X_{12}	overcast	mild	high	strong	yes
X_{13}	overcast	hot	normal	weak	yes
X_{14}	rain	mild	high	strong	no

- The nearest neighbors of X_1 in its class "no" (near hits) are X_2 and X_8 (ignoring the class y we have: $D(X_1, X_2) = 1$, $D(X_1, X_6) = 4$, $D(X_1, X_8) = 1$, $D(X_1, X_{14}) = 3$).
- Attribute x_4 (wind) has different values in X_1 and X_2, so we decrease its relevance.
- Attribute x_2 (temperature) has different values in X_1 and X_8, so we decrease its relevance too.
- The nearest neighbor of X_1 in the opposite class "yes" (near miss) is X_3 ($D(X_1, X_3) = 1$).
- Attribute x_1 (outlook) has different values in X_1 and X_3, so we increase its relevance.

3 Entropy-based attribute selection

- Let S be a set of vectors from m classes – $C_1, C_2, ..., C_m$. Then the number of vectors in S is $|S| = |S_1| + |S_2| + ... + |S_m|$, where S_i is the set of vectors from class C_i.
- The entropy of the class distribution in S (or the average information needed to classify an arbitrary vector) is

  $I(S) = -P(C_1) \times log_2 P(C_1) - P(C_2) \times log_2 P(C_2) - ... - P(C_n) \times log_2 P(C_n)$,

  where $P(C_i) = \frac{|S_i|}{|S|}$.
- Assume that attribute A splits S into k subsets – $A_1, A_2, ..., A_k$ (each A_i having the same value for A).
- Then the information in the split, based on the values of A is

$$I(A) = \frac{|A_1|}{|S|} \times I(A_1) + \frac{|A_2|}{|S|} \times I(A_2) + ... + \frac{|A_k|}{|S|} \times I(A_k))$$

- Then, the *information gain* is

$$gain(A) = I(S) - I(A)$$

- The most *relevant attribute* (the one with the highest discriminant power) is the attribute with *maximal information gain.*

Example

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
X	x_1	x_2	x_3	x_4	y
X_1	sunny	hot	high	weak	no
X_2	sunny	hot	high	strong	no
X_3	overcast	hot	high	weak	yes
X_4	rain	mild	high	weak	yes
X_5	rain	cool	normal	weak	yes
X_6	rain	cool	normal	strong	no
X_7	overcast	cool	normal	strong	yes
X_8	sunny	mild	high	weak	no
X_9	sunny	cool	normal	weak	yes
X_{10}	rain	mild	normal	weak	yes
X_{11}	sunny	mild	normal	strong	yes
X_{12}	overcast	mild	high	strong	yes
X_{13}	overcast	hot	normal	weak	yes
X_{14}	rain	mild	high	strong	no

- $I(S) = -P(yes) \times log_2 P(yes) - P(no) \times log_2 P(no) =$
  $-\frac{5}{14} \times log_2 \frac{5}{14} - \frac{9}{14} \times log_2 \frac{9}{14}$
- $A = outlook$, $A_1 = \{1, 2, 8, 9, 11\}$ (sunny),
  $A_2 = \{3, 7, 12, 13\}$ (overcast),
  $A_3 = \{4, 5, 6, 10, 14\}$ (rainy).
- $I(outlook) = \frac{5}{14} \times I(A_1) + \frac{4}{14} \times I(A_2) + \frac{5}{14} \times I(A_3)$
- $I(A_1) = I(\{no, no, no, yes, yes\}) = -\frac{3}{5} \times log_2 \frac{3}{5} - \frac{2}{5} \times log_2 \frac{2}{5}$
- $I(A_2) = I(\{yes, yes, yes, yes\}) = 0$
- $I(A_3) = I(\{yes, yes, no, yes, no\}) = -\frac{3}{5} \times log_2 \frac{3}{5} - \frac{2}{5} \times log_2 \frac{2}{5}$
- Best attribute $\Rightarrow$ *outlook*

4 Statistical measures

4.1 Bacis measures

- Measuring central tendency
  - *Arithmetic mean* (average) of all values of an attribute:

$$\mu = \frac{1}{n}\sum_{1}^{n} x_i$$

  - *Median*: the middle value in an ordered sequence.
- Measuring *dispersion*: variance (σ) and standard deviation (σ^2)

$$\sigma^2 = \frac{1}{n}\sum_{1}^{n}(x_i - \mu)^2$$

- Measuring probability (density function)

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2}\pi\sigma}\, e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

4.2 Correlation analysis

Correlation between occurrences of A and B:

$$corr(A, B) = \frac{P(A, B)}{P(A)P(B)}$$

- $corr(A, B) < 1 \Rightarrow A$ and B are negatively correlated.
- $corr(A, B) > 1 \Rightarrow A$ and B are positively correlated.
- $corr(A, B) = 1 \Rightarrow A$ and B are independent.

Contingency table (weather data)

	outlook=sunny	outlook≠sunny	Row total
play=yes	2	7	9
play=no	3	2	5
Column total	5	9	14

$$corr(outlook = sunny, play = yes) = \frac{\frac{2}{14}}{\frac{5}{14} \times \frac{9}{14}} = 0.62 < 1$$

$\Rightarrow$ negaive correlation

$$corr(outlook = sunny, play = no) = \frac{\frac{3}{14}}{\frac{5}{14} \times \frac{5}{14}} = 1.68 > 1$$

$\Rightarrow$ positive correlation

4.3 The χ^2 test

Assume term t with values $\{0,1\}$ and class C with values $\{0,1\}$ are two random variables with n observations (e.g. document vectors, where t appears or not). To find out whether t and C are independent or not we use the following test.

$$\chi^2 = \sum_{l,m} \frac{(P(C=l,t=m) - nP(C=l)P(t=m))^2}{nP(C=l)P(t=m)}$$

The higher the value of χ^2, the lower is our belief that these variables are independent given the observed data. We may compute χ^2 using the contingency matrix.

	$t=0$	$t=1$	Row total
$C=0$	k_{00}	k_{01}	$k_{00}+k_{01}$
$C=1$	k_{10}	k_{11}	$k_{10}+k_{11}$
Column total	$k_{00}+k_{10}$	$k_{01}+k_{11}$	n

$$\begin{aligned}\chi^2 = {} & \frac{(k_{00} - n(k_{00}+k_{01})(k_{00}+k_{10}))^2}{n(k_{00}+k_{01})(k_{00}+k_{10})} + \\ & \frac{(k_{01} - n(k_{00}+k_{01})(k_{01}+k_{11}))^2}{n(k_{00}+k_{01})(k_{01}+k_{11})} + \\ & \frac{(k_{10} - n(k_{10}+k_{11})(k_{00}+k_{10}))^2}{n(k_{10}+k_{11})(k_{00}+k_{10})} + \\ & \frac{(k_{11} - n(k_{10}+k_{11})(k_{01}+k_{11}))^2}{n(k_{10}+k_{11})(k_{01}+k_{11})}\end{aligned}$$

For the purposes of feature selection we prefer terms with higher χ^2 values (higher dependence between the term and the class variable). To rank features we order them by their χ^2 values in decreasing order.

4.4 Mutual Information

Assume X and Y are discrete random variable taking values denoted by x and y. The mutual information between X and Y is defined as follows:

$$M(X, Y) = \sum_x \sum_y P(x, y) \log \frac{P(x, y)}{P(x)P(y)}$$

- M is similar to *entropy* (information): $H(X) = \sum_x P(x) \log P(x)$. $M(X, Y)$ is the reduction in the entropy of X if the value of Y is known (and vice versa).
- When X and Y are independent $M(X, Y) = 0$.