Probability

$P(A) = |A|/|\Omega|$

Conditional probability

$P(A|B) = P(A, B)/P(B)$

Statistical independence

$A\ ind.\ B \Leftrightarrow P(A|B) = P(A)(P(A, B) = P(A) \cdot P(B)$

Entropy

$H(X) = \sum_{x \in X} p(x) \cdot log_2(1/p(x))$

Conditional entropy

$H(X|Y) = \sum_{x \in X, y \in Y} p(x, y) \cdot \log_2(p(y)/p(x, y))$
$= H(X, Y) - H(Y) = \sum_{x \in X, y \in Y} p(x, y) \cdot \log_2(1/p(x, y)) - \sum_{y \in Y} p(y) \cdot \log_2(1/p(y))$ $I(H, X) = \sum_{x \in X, y \in Y} p(x, y) \cdot \log_2(p(x, y)/(p(x) \cdot p(y)))$

Evaluation measures, Confusion matrix

- accuracy: $(TP + TN)/total$
- error rate: $(FP + FN)/total$
- precision: $TP/total\ positive$
- sensitivity/recall: $TP/actual\ yes$ (true positive rate)
- specificity: $TN/actual\ no$ (true negativity rate)
- prevalence: $actual\ yes/total$
- Cohen's Kappa: $\frac{p_A - \sum p(,i) \cdot p(i,)}{1 - \sum p(,i) \cdot p(i,)}$
- F score: $F_1 = 2\frac{precision \cdot recall}{precision + recall}$ (can be generalized to F_β)

Inter-rater agreement

$sum\ diagnoal/all$

Statistical data analysis

- expected value of a random variable X: $E[X] = \sum_{x \in X} p(x) \cdot x$
- variance: $\sigma^2 = Var(X) = E[(X - \mu)^2] = \sum p_i \cdot (x_i - \mu)^2, \mu = avg(X)$
  variance of a set $= \frac{1}{n} \sum (x - \mu)^2$
- covariance: σ_{XY}
  $E[(X - \mu)(Y - \mu)] = \sum p_i \cdot (x_i - E[X])(y_i - E[Y])$
  sample covariance: $\rho_{X,Y} = \frac{1}{N-1} \sum_1^N (x_i - \bar{x})(y_i - \bar{y})$
- Pearson correlation coefficient (correlation): $-1 \leq \rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \leq 1$
- standard deviation: $\sigma_X = \sqrt{\sigma_X^2}$
- median: $2 - q(1)$
- quantiles: $k - q_X(m) = X_l : |i : X_i \leq X_l| = \frac{x \cdot |X|}{k} \cdots X_l = X_{\frac{x \cdot |X|}{k}}$

Pearson's χ^2 test

...

Clustering

- partitioning of data set
- centroid: $\mu(C_i) = \frac{1}{|C_i|} \sum_{x \in C_i} x$
- within-cluster variation: $L(C_i) = 2 \sum_{x \in C_i} d(x, \mu(C_i))$ (d is the distance function)
- total within-cluster variation: $L(C_1, \cdots) = \sum L(C_i)$
- optimalization task: $argmin_{C_1, C_2, \ldots} L(C_1, \cdots)$

K-means

1. $C_1^0 = x_{random}, C_2^0 = x_{random}, \cdots$
2. centroid update (compute $\mu(C_i)$)
3. data assignment (assign data to closest centroid)
4. if clusters remain the same, done, else goto 2

Dendrograms

- rooted binary tree
- $height = distance$ (node location at the y axis is the dissimilarity between child groups)
- two methods:
    1. merge two most similar clusters
    2. top down ??
- closest clusters $(C_{n_1}, C_{n_2}) = argmin_{C_u, C_v} d(C_u, C_v)$ (d is the linkage function)
-
    1. single linkage: minimum between clusters ($d(C_i, C_j) = min_{x \in C_i, y \in C_j} d(x, y)$)
    2. complete linkage: maximum between clusters ($d(C_i, C_j) = max_{x \in C_i, y \in C_j} d(x, y)$)
    3. average linkage: avg between all elements in clusters ($d(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i, y \in C_j} d(x, y)$)