CSCI 347 Cheat Sheet: Exploratory Data Analysis

Sample mean of an attribute X_j:

$$\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

Sample variance of X_j :

$$\hat{\sigma}_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \hat{\mu})^2$$

Multivariate/multi-dimensional mean, when x_i is a data instance represented as a d -dimensional vector:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$ (and $\hat{\mu}$ is also a d -dimensional vector)

Sample covariance (between attributes X_p and X_q where there are n data instances):

$$\hat{\sigma}_{pq} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ip} - \hat{\mu}_p)(x_{iq} - \hat{\mu}_q)$$

Covariance matrix:

$$\Sigma = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \hat{\sigma}_{13} \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 & \hat{\sigma}_{23} \\ \hat{\sigma}_{31} & \hat{\sigma}_{32} & \hat{\sigma}_3^2 \end{pmatrix}$$

Correlation:

$$\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1\hat{\sigma}_2}$$

Total variance of data matrix *D* with *d* attributes:

$$\mathrm{Var}(D) = \hat{\sigma}_1^2 + \hat{\sigma}_2^2 + \ldots + \hat{\sigma}_d^2$$

Range normalization:

$$x_i' = \frac{x_i - \min_{i=1}^{n}\{x_i\}}{\max_{i=1}^{n}\{x_i\} - \min_{i=1}^{n}\{x_i\}}$$

Z-score normalization:

$$x_i' = \frac{x_i - \hat{\mu}_j}{\hat{\sigma}_j}$$ for an attribute X_j

L_2 norm of a vector x_i with d dimensions (columns/attributes):

$$||x_i||_2 = \sqrt{\sum_{k=1}^{d} x_{ik}^2}$$

L_1 norm:

$||x_i - x_j||_1 = \sum_{k=1}^{m} |x_{ik} - x_{jk}|$ where x_i and x_j are vectors, and there are m dimensions

Dot product:

$a^T b = \sum_{k=1}^{m} a_k b_k$

where a and b are vectors, and there are m dimensions

Cosine similarity (cosine of the angle between two vectors x_i and x_j) :

$cos(\theta) = \dfrac{x_i^T x_j}{||x_i||_2 ||x_j||_2}$ where x_i and x_j are vectors and $x_i^T x_j$ is their dot product

Label encoding: assigns each of *k* possible values for a categorical variable a unique integer between 1 and k (scikit-learn's LabelEncoder implementation assigns an integer between 0 and k-1)

One-hot encoding: converts each categorical variable with *k* possible values into a *k*-dimensional vector with a one-of-k encoding scheme. I.e., a categorical variable with *k* categories is encoded as a a *k*-dimensional binary vector with *k-1* zeros and one 1.

Hamming Distance: $\delta_H(x_i, x_j)$ = number of entries where x_i and x_j do not have the same value. When x_i and x_j consist of categorical attributes that have been one-hot-encoded, $\delta_H(x_i, x_j) = d - s$, where d is the number of categorical attributes, and s is the number of those attributes that match in value in x_i and x_j.

Jaccard similarity:

The ratio of the number of matching values to the number of distinct values that appear in both data instances

$J(x_i, x_j) = \dfrac{s}{2(d-s)+s} = \dfrac{s}{2d-s}$ where *s* and *d* are defined as above.

Gower distance:

$$G(x_i, x_j) = \frac{1}{d}\sum_{k=1}^{d} \text{dist}_{ij}(k)$$

where for categorical attributes,

$$\text{dist}_{ij}(k) = \begin{cases} 0 & \text{if } x_{ik} = x_{jk} \\ 1 & \text{otherwise} \end{cases}$$

and for numerical attributes,

$$\text{dist}_{ij}(k) = \frac{|x_{ik} - x_{jk}|}{\text{Range}(k)}$$