Statistics

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

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

Distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$P(x) = (1-p)^{k-1}\,p$$

$$P(x) = p^x(1-p)^{1-x} \qquad f(x) = \frac{1}{n} \qquad f(x) = \frac{1}{b-a}$$

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad P\{c \le X \le d\} = \frac{d-c}{b-a}$$

Association

$$Z_{stat} = \frac{\bar{x}-\mu_0}{SE_{\bar{x}}} \qquad SE_x = \frac{\sigma}{\sqrt{n}}$$

$$k = \frac{P_o - P_e}{1-P_e} \qquad P_e = \frac{a+d}{N} \qquad P_o = \frac{(a+c)(a+b)+(b+d)(c+d)}{N^2} \qquad r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \qquad r_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\cdot\sqrt{\sum(y_i-\bar{y})^2}}$$

$$y = a \cdot x + b \qquad a = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} \qquad b = \bar{Y} - a\bar{X} \qquad SSE = \sum(y_i - \hat{y}_i)^2$$

Preprocessing

$$v' = \frac{v-min}{max-min}(new\_max - new\_min) + new\_min \qquad v' = \frac{v-\mu}{\sigma} \qquad v' = \frac{v}{10^j} \qquad W = (B-A)/N$$

Machine Learning

$$\text{Precision} = \frac{TP}{TP+FP} \Big/ \frac{TN}{TN+FN}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad \text{Sensitivity}$$

$$F_\beta = \frac{(1+\beta^2)\cdot P \cdot R}{\beta^2 \cdot P + R} \qquad F_1 = \frac{2 \cdot P \cdot R}{P+R} \qquad \frac{TN}{TN+FP} \quad \text{Specifity}$$

	+	−
+	a TP	b FP
−	c FN	d TN

Time series analysis

$$\text{ACF: } r_k = \frac{\sum(x_t-\bar{x})(x_{t-k}-\bar{x})}{\sum(x_t-\bar{x})^2}$$

Models: $y_t = T_t + S_t + R_t$, $y_t = T_t \cdot S_t \cdot R_t$

$$\text{AR: } x_t = c + \sum_{i=1}^{p}\varphi_i x_{t-i} + \varepsilon_t$$

$$\text{MA: } x_t = \mu + \varepsilon_t + \sum_{i=1}^{q}\theta_i \varepsilon_{t-i}$$

$$AIC = T \cdot \log\left(\frac{SSE}{T}\right) + 2(k+2)$$

$$\text{Non-Seasonal ARMA: } x'_t = c + \sum_{i=1}^{p}\varphi_i x'_{t-i} + \sum_{j=1}^{q}\theta_j \varepsilon_{t-j} + \varepsilon_t$$

$$AIC_c = AIC + \frac{2(k+2)(k+3)}{T-k-3}$$

$$\text{Ljung-Test: } Q = T(T+2)\sum_{k=1}^{h}(T-k)^{-1} r_k^2$$

Spatio-temporal data analysis

$$\hat{C}_z^{(\tau)}(s_i; s_k) \equiv \frac{1}{T-\tau}\sum_{j=\tau+1}^{T}\left(Z(s_i;t_j) - \hat{\mu}_{z,s}(s_i)\right)\left(Z(s_k;t_j-\tau) - \hat{\mu}_{z,s}(s_k)\right)$$

$$\text{Spatial: } \hat{\mu}_{z,s}(s_i) = \frac{1}{T}\sum_{i=1}^{T} Z(s_i;t_i) \qquad \text{Temporal: } \hat{\mu}_{z,s}(t_j) = \frac{1}{m}\sum_{i=1}^{m} Z(s_i;t_j)$$

Covariogram:

$$\hat{C}_z(h;\tau) = \frac{1}{|N_s(h)|}\frac{1}{|N_t(\tau)|}\sum_{s_i s_j \in N_s(h); t_j, t_l \in N_t(\tau)} \sum \left(Z(s_i;t_j) - \hat{\mu}_{z,s}(s_i)\right)\left(Z(s_k;t_j) - \hat{\mu}_{z,s}(s_k)\right)$$

Semivariogram:

$$\hat{\gamma}_z(h;\tau) = \frac{1}{|N_s(h)|}\frac{1}{|N_t(\tau)|}\sum_{s_i s_j \in N_s(h); t_j, t_l \in N_t(\tau)} \sum \left(Z(s_i;t_j) - Z(s_k;t_l)\right)^2$$

$N_s(h)$: pairs with spatial lag in h

$N_t(\tau)$: pairs with time lag in τ

$$\hat{Z}(s_0;t_0) = \sum_{j=1}^{T}\sum_{i=1}^{m_j} w_{ij}(s_0;t_0)\,Z(s_{ij};t_j) \qquad w_{ij}(s_0;t_0) = \frac{\tilde{w}_{ij}(s_0;t_0)}{\sum_{k=1}^{T}\sum_{l=1}^{m_k}\tilde{w}_{lk}(s_0;t_0)} \qquad \tilde{w}_{ij}(s_0;t_0) = \frac{1}{d((s_{ij};t_j),(s_0;t_0))^\alpha}$$

Text data analysis

$$w_{t,d} = tf_{t,d} \cdot idf_t \qquad tf_{t,d} = \begin{cases} 1 + \log_{10}(count(t,d)) & count > 0 \\ 0 & else \end{cases} \qquad idf_t = \log_{10}\left(\frac{N}{df_t}\right)$$

$$PMI(w,c) = \log_2\left(\frac{P(w,c)}{P(w)P(c)}\right) \qquad P(X,Y) = P(X)P(Y|X) \qquad PPMI(w,c) = \max(PMI, 0)$$

$$P(w_i,c_j) = \frac{f_{ij}}{\sum_{n\in W}\sum_{m\in C} f_{nm}} \qquad P(w_i) = \frac{\sum_{m\in C} f_{im}}{\sum_{n\in W}\sum_{m\in C} f_{nm}} \qquad P(c_j) = \frac{\sum_{n\in W} f_{nj}}{\sum_{n\in W}\sum_{m\in C} f_{nm}} \qquad \cos(\vec{u},\vec{v}) = \frac{\vec{u}^T \cdot \vec{v}}{\|\vec{u}\| \cdot \|\vec{v}\|}$$