General
Variance: $\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{N-1}$
Standard Deviation: $\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N-1}}$
Pearson: $\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{N}\sum (x_i-\bar{x})(y_i-\bar{y})}{\sigma_X\sigma_Y}, -1 < \rho < 1$
Spearman: $1 - \frac{6\Sigma d^2}{n(n^2-1)}, -1 < \rho < 1$
Point Biserial: $r_{\mathrm{pb}} = \frac{M_1 - M_0}{\sigma_Y}\sqrt{p_1 p_2}$
ANOVA: $F = \frac{\mathrm{var}_b}{\mathrm{var}_w} = \frac{\mathrm{MSB}}{\mathrm{MSW}} = \frac{\frac{\mathrm{SS_b}}{df_b}}{\frac{\mathrm{SS_w}}{df_w}}, df_{\mathrm{b}} = \mathrm{classes} - 1, df_w = m - \mathrm{classes}$
Chi-squared: $\chi^2 = \sum \frac{(O-E)^2}{E}, E = \frac{O_A \times O_b}{O_{\mathrm{total}}}$
Min-Max Scaling: $X' = \frac{X - \min(X)}{\max(X) - \min(X)}$
Standard Scaling: $\mu = 0, \sigma = 1$: $Z = \frac{X-\mu}{\sigma}$

Linear Regression:
Cost: $J(w) = \frac{1}{2m}\sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)})^2 = \frac{1}{2m}\sum_{i=1}^{m} (w^\top x^{(i)} - y^{(i)})^2$
Derivative: $\frac{\partial J(w)}{\partial w} = \frac{1}{m}\sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)}) \cdot x_n^{(i)}$
Update: $w_n := w_n - \alpha \frac{\partial J(w)}{\partial w} = w_n - \alpha \frac{1}{m}\sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)}) \cdot x_n^{(i)}$
Normal equation: $J(w) = \frac{1}{2m}(w^\top X^\top X w - 2w^\top X^\top y - y^\top y)$
Derivative for normal equation: $\frac{\partial J(w)}{\partial w} = \frac{1}{m}(X^\top X w - X^\top y)$
Parameters: $w = (X^\top X)^{-1} X^\top y$
OLS (w_0): $w_1 = \frac{\sum_{i=1}^{m} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{m} (x_i - \bar{x})^2}$
OLS (w_1): $w_0 = \bar{y} - w_1 \bar{x}$

Metrics:
RMSE: $\sqrt{\frac{1}{m}\sum_{i=1}^{m} (y^{(i)} - \hat{y}^{(i)})^2}, 0 \to \infty$
$R^2 = 1 - \frac{\mathrm{SS_{res}}}{\mathrm{SS_{tot}}}$
$\mathrm{SS_{tot}} = \sum_{i=1}^{m} (y^{(i)} - \bar{y}^{(i)})^2$
$\mathrm{SS_{reg}} = \sum_{i=1}^{m} (\hat{y}^{(i)} - \bar{y}^{(i)})^2$
$\mathrm{SS_{res}} = \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)})^2$
Adjusted R^2: $R^2_{\mathrm{adj}} = 1 - \frac{(1-R^2)(m-1)}{m-n-1}$
Precision: $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$
Recall: $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$
F1: $2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$

Logistic Regression
$h_w(x) = g(w^\top x)$, where $g(w^\top x) = \frac{1}{1+e^{-(w^\top x)}}$
Cost: $J(w) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)}\log(h_w(x^{(i)})) + (1-y^{(i)})\log(1-h_w(x^{(i)}))$
$\frac{\partial h_w(x)}{\partial w_j} = h_w(x)(1-h_w(x))x_j$
Derivative: $\frac{\partial}{\partial w_j} J(w) = \frac{1}{m}\sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)}) x_j^{(i)}$
Update: $w_j := w_j - \alpha \frac{1}{m}\sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)}) x_j^{(i)}$
Odds: $\frac{P}{1-P}$, $\mathrm{logit}(P) = \log\left(\frac{P}{1-P}\right)$
Coefficients: $\mathrm{logit} = w^\top x$, $\mathrm{odds} = e^{w^\top x}$

Ridge (Linear Regression)
Cost: $J(w) = \frac{1}{2m}\left[\sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} w_j^2\right]$
Derivative: $\frac{\partial J(w)}{\partial w} = \frac{1}{m}\sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)}) + \frac{\lambda}{m} w_j$
Update: $w_0 := w_0 - \frac{\alpha}{m}\sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)})$
Update: $w_j := w_j\left(1 - \alpha\frac{\lambda}{m}\right) - \frac{\alpha}{m}\sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)}) x_j$
Normal Equation: $w = \left(X^\top X + \lambda \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix}\right)^{-1} X^\top y, I \in \mathbb{R}^{n-1 \times n-1}$

LASSO (Linear Regression)
Cost: $J(w) = \frac{1}{2m}\left[\sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} |w_j|\right]$
Derivative: $\frac{\partial J(w)}{\partial w} = \frac{1}{m}\sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)}) + \frac{\lambda}{2m}\mathrm{sign}(w_j)$
Update: $w_0 := w_0 - \frac{\alpha}{m}\sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)})$
Update: $w_j := w_j - \alpha\frac{\lambda}{m} \cdot \mathrm{sign}(w_j) - \frac{\alpha}{m}\sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)}) x_j$

L2 (Logistic Regression)
Cost: $J(w) = \left[-\frac{1}{m}\sum_{i=1}^{m} y^{(i)}\log(h_w(x^{(i)})) + (1-y^{(i)})\log(1-h_w(x^{(i)}))\right] + \frac{\lambda}{2m}\sum_{j=1}^{n} w_j^2$
Derivative: $\frac{\partial J(w)}{\partial w} = \frac{1}{m}\sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)}) + \frac{\lambda}{m} w_j$
Update: $w_0 := w_0 - \frac{\alpha}{m}\sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)})$
Update: $w_j := w_j\left(1 - \alpha\frac{\lambda}{m}\right) - \frac{\alpha}{m}\sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)}) x_j$

L1 (Logistic Regression)
Cost: $J(w) = \left[-\frac{1}{m}\sum_{i=1}^{m} y^{(i)}\log(h_w(x^{(i)})) + (1-y^{(i)})\log(1-h_w(x^{(i)}))\right] + \frac{\lambda}{2m}\sum_{j=1}^{n} |w_j|$
Derivative: $\frac{\partial J(w)}{\partial w} = \frac{1}{m}\sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)}) + \frac{\lambda}{2m}\mathrm{sign}(w_j)$
Update: $w_0 := w_0 - \frac{\alpha}{m}\sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)})$
Update: $w_j := w_j - \alpha\frac{\lambda}{m} \cdot \mathrm{sign}(w_j) - \frac{\alpha}{m}\sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)}) x_j$

Decision Trees
Entropy: $H(p_1) = -\sum_{i=1}^{n} p_i \log_2(p_i)$
Gini: $\sum_k p_k(1-p_k) = 1 - \sum_k p_k^2$
Information gain: $H(p_1^{\mathrm{root}}) - \left(w^{\mathrm{left}} H(p_1^{\mathrm{left}}) + w^{\mathrm{right}} H\left(p_1^{\mathrm{right}}\right)\right)$

SVMs
Soft: $\min_w C\sum_{i=1}^{m} \max(0, 1 - y^{(i)}(w^\top x^{(i)})) + \frac{1}{2}\sum_{j=1}^{n} w_j^2, C = \frac{1}{\lambda}$
Hard: $\min_w \frac{1}{2}\|w\|^2$
Kernel: $\min_w C\sum_{i=1}^{m} \max(0, 1 - y^{(i)}(w^\top f^{(i)})) + \frac{1}{2}\sum_{j=1}^{m} w_j^2$
$f = K(x, l^{(i)})$

Anomaly Detection
Gaussian Distribution: $p(x_j; \mu_j, \sigma_j^2) = \frac{1}{\sqrt{2\pi}\sigma_j}\exp\left(-\frac{(x_j-\mu_j)^2}{2\sigma_j^2}\right)$
Density Estimation: $p(x) = \prod_{j=1}^{n} p(x_j; \mu_j, \sigma_j^2) < \varepsilon$
Multivariate: $p(x_1; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$
$\Sigma = \frac{1}{m}\sum_{i=1}^{m} (x^{(i)} - \mu)(x^{(i)} - \mu)^\top, \mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$

Feature Selection / Dimensionality Reduction
Cramer's V: $\sqrt{\frac{\chi^2}{n(k-1)}}$
Correlation Ratio: $\eta^2 = \frac{\mathrm{SS_{between}}}{\mathrm{SS_{total}}}, \to \eta = \sqrt{\eta^2}, 0 < \eta < 1$
Covariance Matrix: $\Sigma = \frac{1}{m}\sum_{i=1}^{m} (x^{(i)})(x^{(i)})^\top$, find λ
TSNE:

$$p_{j|i} = \frac{\exp\left(-\frac{\|x^{(i)} - x^{(j)}\|^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(-\frac{\|x^{(i)} - x^{(k)}\|^2}{2\sigma_i^2}\right)}, p_{ij} = \frac{p_{j|i} + p_{i|j}}{2m}$$

Perplexity: $\mathscr{P} = 2^{\mathscr{H}}$, $\mathscr{H} = -\sum_{j \neq i} p_{j|i}\log_2(p_{j|i})$
$q_{ij} = \frac{w_{ij}}{Z}, w_{ij} = k(\|y_i - y_j\|), Z = \sum_{k \neq l} w_{kl}$

Similarity kernel in SNE: $k(d) = \exp(-d^2)$
Similarity kernel in tSNE: $k(d) = \frac{1}{1+d^2}$
$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_i - y_j\|^2)^{-1}}$

Cost function (Kullback-Leibler Divergence) :

$$J = \sum_i \mathrm{KL}(p_{ij} | q_{ij}) = \sum_{i \neq j} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right)$$

$$= \sum_{i \neq j} p_{ij}\log(p_{ij}) - \sum_{i \neq j} p_{ij}\log(w_{ij}) + \log\sum_{i \neq j} (w_{ij})$$

Update rule:

$$\frac{\partial}{\partial y_i} J = 4\sum_{j \neq i} (p_{ij} - q_{ij}) q_{ij} (y_i - y_j)$$