FreeFellow

FORMULA SHEET

Exam PA

SOA · Predictive Analytics

53
FORMULAS

5
TOPICS

freefellow.org/soa-pa/formulas

PREDICTIVE ANALYTICS PROBLEM DEFINITION

7 items

Profit per policy

$\text{Profit} = P - L - E$, P = premium (price) per policy, L = expected loss (claim severity times frequency), E = expense or acquisition cost per policy

Value as price minus cost

$V = P - C$, V = value of the item, P = price (e.g. sale price per building), C = cost (e.g. construction cost per building)

Base rate of a binary target

$p = \frac{n_{event}}{N}$, p = base rate (positive-class proportion), n_event = number of positive-class records, N = total records

Majority-class (no-information) accuracy

$acc_{maj} = 1 - p$, acc_maj = accuracy of always predicting the majority class, p = base rate of the rare (positive) class

Business impact of a predictive model

$\text{Impact} = V \times f$, V = dollar value of the decision the prediction improves, f = how often that decision is made

Additive error data-generating model

$y = f(x) + \varepsilon$, y = observed response, f(x) = true signal function, ε = random noise with variance $\text{Var}(\varepsilon)$

Bias-variance decomposition of expected test MSE

$E\left[(y_0 - \hat{f}(x_0))^2\right] = \text{Var}(\hat{f}(x_0)) + \left[\text{Bias}(\hat{f}(x_0))\right]^2 + \text{Var}(\varepsilon)$, $\hat{f}$ = fitted model, x_0 = test point, y_0 = true response, $\text{Var}(\varepsilon)$ = irreducible error

DATA EXPLORATION AND VISUALIZATION

8 items

Variance of a factor-level estimate

$\text{Var}(\hat{\mu}_{\text{level}}) \propto \frac{1}{n_{\text{level}}}$, $\hat{\mu}_{\text{level}}$ = estimated mean for that level, n_{level} = number of rows in that level

Minority rows needed to oversample to a target share

$f = \frac{pN}{1-p}$ from $p = \frac{f}{N+f}$, p = target minority proportion, N = majority-class row count, f = minority rows after oversampling

Boxplot outlier fences

$\text{lower} = Q1 - 1.5 \times \text{IQR},\ \text{upper} = Q3 + 1.5 \times \text{IQR}$, Q1 = first quartile, Q3 = third quartile, IQR = interquartile range

Interquartile range

$\text{IQR} = Q3 - Q1$, Q1 = 25th percentile (first quartile), Q3 = 75th percentile (third quartile)

Shared variance between two predictors

$r^2 = r_{xy}^2$, r_{xy} = Pearson correlation between the two variables; r^2 = proportion of one variable's variation explained by the other

Sample variance

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, x_i = observation, $\bar{x}$ = sample mean, n = sample size

Pearson correlation coefficient

$r_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}}$, x,y = paired values, $\bar{x}$, $\bar{y}$ = means, i = observation index

Sample standard deviation

$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$, s^2 = sample variance, x_i = observation, $\bar{x}$ = mean, n = sample size

DATA TRANSFORMATIONS AND UNSUPERVISED LEARNING TECHNIQUES

12 items

Standardized feature (z-score for clustering)

$z = \frac{x - \mu}{\sigma}$, x = raw feature value, μ = feature mean, σ = feature standard deviation

Complete linkage between-cluster distance

$d(A, B) = \max_{a \in A,\, b \in B} \|a - b\|$, A, B = clusters, a, b = member points, distance = farthest pair

Proportion of variance explained by a principal component

$\mathrm{PVE}_m = \frac{\mathrm{Var}(z_m)}{\sum_j \mathrm{Var}(z_j)}$, PVEm = proportion for PC m, Var(zm) = variance of PC m, denominator = total variance across all PCs

Variance of a principal component from its standard deviation

$\mathrm{Var}(z_m) = s_m^2$, Var(zm) = variance captured by PC m, sm = reported standard deviation of PC m; total variance = number of standardized variables

Within-cluster sum of squares (K-means objective)

$\mathrm{WCSS} = \sum_{k=1}^{K} \sum_{x \in C_k} \|x - \mu_k\|^2$, K = number of clusters, C_k = cluster k, x = observation, μ_k = mean vector of cluster k

Single linkage between-cluster distance

$d(A, B) = \min_{a \in A,\, b \in B} \|a - b\|$, A, B = clusters, a, b = member points, distance = closest pair

First principal component score

$z_1 = \phi_{11} x_1 + \phi_{21} x_2 + \phi_{31} x_3$, z1 = PC1 score, φj1 = loading of variable j on PC1, xj = standardized feature j

Loading vector normalization constraint

$\sum_j \phi_{jm}^2 = 1$, φjm = loading of variable j on PC m; squared loadings of a component sum to one (unit-length loading vector)

Number of dummy columns for a categorical variable

$d = k - 1$, k = number of category levels, d = number of dummy (indicator) columns needed to avoid redundancy in a GLM

Log transform for a skewed non-negative variable

$x' = \ln(x + 1)$, x = raw strictly non-negative value, x' = transformed value; the +1 lets x = 0 be handled

Min-max scaling

$x_{\mathrm{mm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$, x = raw value, $x_{\min}$ = minimum, $x_{\max}$ = maximum, x_{mm} = scaled value in [0,1]

Z-score standardization

$z = \frac{x - \bar{x}}{s}$, x = raw value, $\bar{x}$ = sample mean, s = sample standard deviation, z = standardized value

GENERALIZED LINEAR MODELS

15 items

Poisson log-link model with a log-exposure offset

$\log(E[\text{count}]) = \log(\text{exposure}) + \beta_0 + \beta_1 x_1 + \cdots$, exposure = units of exposure (offset, coef fixed at 1), β_0 = intercept, β_j = slope, x_j = predictor

Weighted least squares objective

$\min_\beta \sum_i w_i (y_i - \hat{y}_i)^2$, w_i = weight of row i, y_i = observed response, $\hat{y}_i$ = fitted value, β = coefficients

Group-size weighted mean of grouped rates

$\bar{y}_w = \dfrac{\sum_i n_i y_i}{\sum_i n_i}$, n_i = group size (weight) of row i, y_i = group-average rate for row i

Expected count as exposure times a fitted rate

$E[\text{count}] = \text{exposure} \times e^{\beta_0 + \beta_1 x_1}$, exposure = units of exposure, $e^{\beta_0 + \beta_1 x_1}$ = rate per unit exposure, β_0 = intercept, β_1 = slope, x_1 = predictor

GLM log link function

$\log(\mu) = \beta_0 + \beta_1 x_1 + \ldots$, so e^{β_1} is the factor by which a one-unit rise in x_1 scales μ; μ = expected response, β = coefficients

Pearson residual for a GLM

$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$, y_i = observed, $\hat{\mu}_i$ = fitted mean, $V(\hat{\mu}_i)$ = variance function at the fitted mean

Lognormal mean back-transformation (smearing correction)

$\hat{E}[Y] = e^{\hat{\mu}_{\log}} e^{\sigma^2/2}$, $\hat{\mu}_{\log}$ = fitted value on log scale, σ^2 = residual variance on log scale

Estimated dispersion parameter

$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$, n = observations, p = parameters, y_i = observed, $\hat{\mu}_i$ = fitted mean, V = variance function

Elastic net penalized regression objective

$\min_\beta \frac{1}{2n} \sum_i (y_i - \beta_0 - \sum_j \beta_j x_{ij})^2 + \lambda [\frac{1-\alpha}{2} \sum_j \beta_j^2 + \alpha \sum_j |\beta_j|]$, λ = penalty strength, α = L1/L2 mix, β = coefficients

Variance-inflation factor for a predictor

$\text{VIF}_j = \frac{1}{1 - R_j^2}$, R_j^2 = R-squared from regressing predictor j on all other predictors

One-standard-error threshold for lambda.1se selection

$\text{threshold} = \text{MSE}_{\min} + \text{SE}_{\min}$, MSE_min = minimum CV mean squared error, SE_min = its standard error; lambda.1se = largest λ with CV MSE ≤ threshold

Log-link multiplicative effect of a coefficient

$\mu_{new} = \mu \times e^{\beta}$ per unit, μ = expected response, β = coefficient on the link scale; $e^{\beta} - 1$ is the percent change

Logit-link odds ratio from a coefficient

$OR = e^{\beta}$, OR = multiplicative change in the odds of the event per unit increase, β = logit-link coefficient

Interaction-adjusted slope of a predictor at a level

$\text{slope of } x_1 \text{ at level } \ell = \beta_1 + \beta_{\text{int},\ell}$, β1 = main coefficient, β_int,l = interaction coefficient (zero for baseline level)

GLM linear predictor with link function

$g(\mu) = \eta = \beta_0 + \beta_1 x_1 + \ldots$, g = link function, μ = expected response, η = linear predictor, β = coefficients, x = predictors

TREE-BASED MODELS

11 items

Default mtry for random forest classification

$m_{try} = \sqrt{p}$, p = total number of predictors, m_try = predictors sampled at each split

Out-of-bag probability a row is left out of a bootstrap sample

$P = (1 - 1/n)^n \to e^{-1} \approx 0.368$, n = number of training rows, P = chance a given row is omitted from one bootstrap (~37%)

Default mtry for random forest regression

$m_{try} = p/3$, p = total number of predictors, m_try = predictors sampled at each split

Gradient boosting additive update

$F_m(x) = F_{m-1}(x) + \lambda h_m(x)$, F = ensemble prediction, m = boosting round, λ = learning rate (shrinkage), h_m = tree fit to current pseudo-residuals

k-fold cross-validation error

$\text{CV error} = \frac{1}{k} \sum_{i=1}^{k} \text{Error}(\text{fold}_i)$, k = number of folds, fold_i = i-th held-out validation fold

Variance of an averaged tree ensemble

$\text{Var} = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$, ρ = pairwise tree correlation, σ² = single-tree prediction variance, B = number of trees

Learning-rate to tree-count inverse tradeoff

$n_{new} \approx n_{old} \times \frac{\lambda_{old}}{\lambda_{new}}$, n = number of trees, λ = learning rate; halving λ roughly doubles the trees needed

Cost-complexity pruning criterion

$R_{cp}(T) = R(T) + cp \cdot |T|$, R(T) = tree total error, cp = complexity parameter, |T| = number of terminal nodes (leaves)

Gini impurity of a node

$G = 1 - \sum_k p_k^2$, G = Gini impurity, p_k = proportion of class k in the node; G = 0 at a pure node (rpart default)

Entropy of a node

$H = -\sum_k p_k \log_2 p_k$, H = entropy in bits, p_k = proportion of class k in the node; H = 0 at a pure node

Root mean square error

$\text{RMSE} = \sqrt{\frac{1}{n} \sum_i (y_i - \hat{y}_i)^2}$, y_i = actual value, $\hat{y}_i$ = predicted value, n = number of observations