$$SD(x) = \sqrt{\mathrm{Var}(x)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i - \overline{\mathbf{x}}\right)^2}.$$

$$Var(\widehat{f(x)}) = E[(\widehat{f(x)} - E[\widehat{f(x)}])^2]$$

$$\mathrm{Corr}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}},$$

Multi-Regression

$$\begin{bmatrix}\hat{y}_1\\ \hat{y}_2\\ \hat{y}_3\\ \vdots\\ \hat{y}_n\end{bmatrix} = \begin{bmatrix}1 & x_{11} & x_{12} & \dots & x_{1p}\\ 1 & x_{21} & x_{22} & \dots & x_{2p}\\ 1 & x_{31} & x_{32} & \dots & x_{3p}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 1 & x_{n1} & x_{n2} & \dots & x_{np}\end{bmatrix}\begin{bmatrix}\theta_0\\ \theta_1\\ \vdots\\ \theta_p\end{bmatrix}$$

L2

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2.$$

L1

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - (b_0 + b_1 x_i)|.$$

Closed Form

$$\widehat{\beta} = (\widetilde{X}^T\,\widetilde{X})^{-1}\,\widetilde{X}^T y$$

$$b_0 = \bar{y} - b_1\bar{x}$$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Sample Covariance

$$Cov(x,y) = \frac{\Sigma(x-\bar{x})(y-y)}{N-1}$$

MAD = median(|preds - true|)

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

$$Correlation = \frac{Cov\,(x,y)}{\sigma x * \sigma y}$$

Correlation Coefficient Formula

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

Logistic

$$p = \frac{1}{1 + e^{-(b_0 + b_1 x)}}$$

Population Covariance Formula

$$Cov(x,y) = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{N}$$

Sample Covariance

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

Logit

$$\log\left(\frac{p}{1-p}\right)$$

odds ^ in parenth

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right)$$

$$\sum_{i \text{ in pos class}}(-\log p_i) + \sum_{i \text{ in neg class}}(-\log(1 - p_i)).$$

$$-\frac{1}{N}\sum_{=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i))$$

Lasso

$$\sum_i(\text{observed price}_i - (b_0 + b_1\text{area}_i))^2 + \lambda(|b_0| + |b_1|)$$

Ridge

$$\sum_i(\text{observed price}_i - (b_0 + b_1\text{area}_i))^2 + \lambda(b_0^2 + b_1^2)$$

Decision Tree

$$\text{Variance measure for split} = \frac{n_{\text{left}}}{n_{\text{parent}}}\text{Var}_{\text{left}} + \frac{n_{\text{right}}}{n_{\text{parent}}}\text{Var}_{\text{right}}$$

$$\text{Gini impurity for split} = \frac{n_{\text{left}}}{n_{\text{parent}}}\text{Gini}_{\text{left}} + \frac{n_{\text{right}}}{n_{\text{parent}}}\text{Gini}_{\text{right}}$$

$$\text{Gini impurity} = 1 - p_1^2 - p_0^2 = 2p_0p_1$$

$$SSE_{L_2} = \sum^{n}(y_i - \hat{y}_i)^2 + \lambda\sum^{P}\beta_j^2$$

$$precision = \frac{TP}{TP + FP}$$

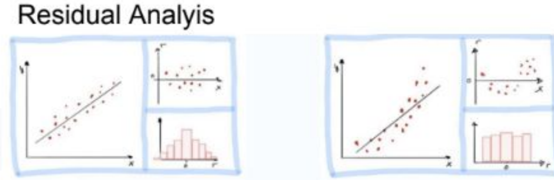
$$recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

$$specificity = \frac{TN}{TN + FP}$$

Residual Analyis



Linear assumption is correct. There is no obvious relationship between residuals and x. Histogram of residuals is symmetric and normally distributed.

Linear assumption is incorrect. There is an obvious relationship between residuals and x. Histogram of residuals is symmetric but not normally distributed.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

sensitivity, recall, hit rate, or true positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = 1 - \mathrm{FNR}$$

specificity, selectivity or true negative rate (TNR)

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{N}} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} = 1 - \mathrm{FPR}$$

Camberra :

$$d(x, y) = \sum_{i=1}^{m}\frac{|x_i - y_i|}{|x_i + y_i|} \qquad (2)$$

Minkowsky :

$$d(x, y) = \left(\sum_{i=1}^{m}|x_i - y_i|^r\right)^{1/r} \qquad (3)$$

Chebychev :

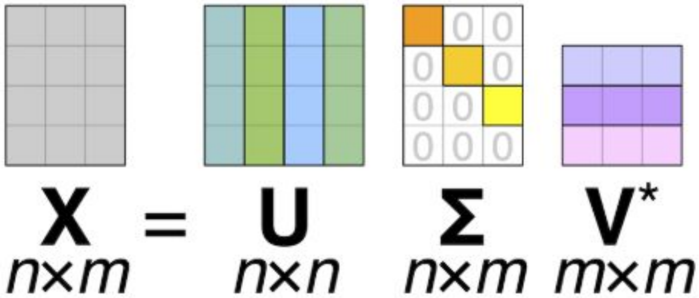
$$d(x, y) = \max_{i=1}^{m}|x_i - y_i| \qquad (4)$$

Euclidean :

$$d(x, y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2} \qquad (5)$$

Manhattan / city - block :

$$d(x, y) = \sum_{i=1}^{m}|x_i - y_i| \qquad (6)$$



$$\sum_j Var(X_j) = \sum_j d_j^2,$$

$$X^{PC} = XV$$

$$\text{prop of variability explained by PC}_j = \frac{d_j^2}{\sum_i d_i^2}$$

K-mans algo

- Given observations $(x_1, x_2, \ldots, x_n)$, partition the observations into k sets $C_1, C_2 \ldots C_k$ to minimize the total within cluster sum-of-squares distances:

$$\text{argmin}_S \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2.$$

Rand Index = # pairs in agreement/ total # of pairs

$$WSS = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} (x_{i,j} - c_{k,j})^2$$

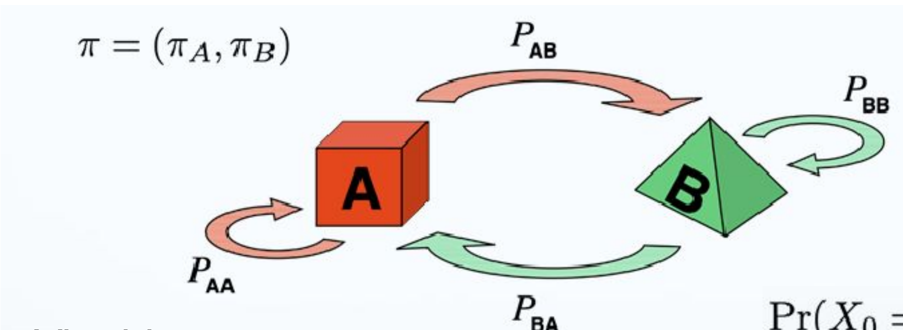
For a single data point:

Let a_i be average distance from point i to other points in same cluster

Let b_i be average distance from point i to points in nearest cluster

Silhouette score:

$$\text{silhouette score}_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

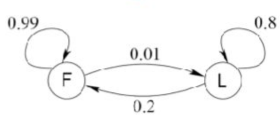


$$\Pr(X_0 = B, X_1 = A, X_2 = A) = \pi_B \cdot p_{BA} \cdot p_{AA}$$

Viterbi

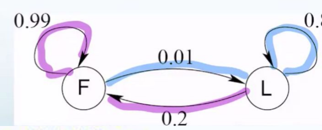
π \ x	ε	6	2	6
B	1	0	0	0
F	0	(1/6)×(1/2) = 1/12	(1/6)×max{(1/12)×0.99, (1/4)×0.2} = 0.01375	(1/6)×max{0.01375×0.99, 0.02×0.2} = 0.00226875
L	0	(1/2)×(1/2) = 1/4	(1/10)×max{(1/12)×0.01, (1/4)×0.8} = 0.02	(1/2)×max{0.01375×0.01, 0.02×0.8} = 0.08

$$v_k(i) = e_k(x_i) \max_r (v_r(i-1) a_{rk})$$



Forward

π	ε	6	2 probability	6 probability of 6
B	1	0	0	0
F	0	(1/6)×(1/2) = 1/12	= 1/6 [1/12 * .99 + 1/4 * .2] = .022	= 1/6 [.022 * .99 + .02 * .2] .0043
L	0	(1/2)×(1/2) = 1/4	= 1/10 [1/12 * .01 + 1/4 * .8) = .02	



Max-Log-Likelihood

$$\log Pr(x|\pi, \mu, \sigma) = \sum_{j=1}^{n} \log(\sum_{i=1}^{k} \pi N(x|\mu_i, \Sigma_i))$$

The most likely path π^* satisfies

Viterbi

$$P(X_t \mid o_{1:T}) = P(X_t \mid o_{1:t}, o_{t+1:T}) \propto P(o_{t+1:T} \mid X_t) P(X_t | o_{1:t})$$

$$\pi^* = \underset{\pi}{argmax}\, Pr(x, \pi)$$

Forward-Backward Algo ^

To find π^*, consider all possible ways the last symbol of x could have been emitted

Gaussian Mixture Model

$$p(\mathbf{x}) = \sum_{i=1}^{k} \pi_i \mathcal{N}(\mathbf{x}|\mu_i, \Sigma_i).$$

Let

$v_k(i)$ = Prob. of path $\langle \pi_1, \cdots, \pi_i \rangle$ most likely to emit $\langle x_1, \ldots, x_i \rangle$ such that $\pi_i = k$

Then

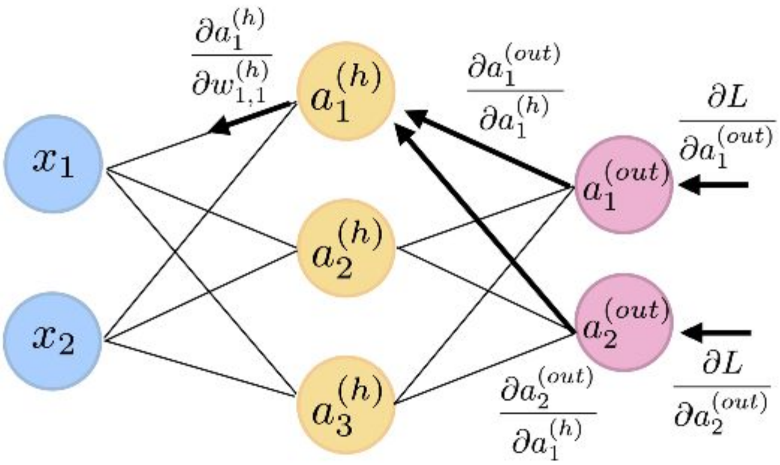
$$v_k(i) = e_k(x_i) \max_r (v_r(i-1) a_{rk})$$

Expectation Maximization

$$Q(\alpha|\alpha^T) = E_{Z|Y,\alpha^T}[\log Pr(Y, Z|\alpha)]$$

$$Q(\alpha|\alpha^T) = \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{y_{k,n} \alpha_k^{(t)}}{\sum_{l=1}^{K} y_{ln} \alpha_l^{(t)}} \log(y_{k,n} \alpha_k)$$

-> Find params that maximize observed data



$$w^{new} = w^{old} - \eta \Delta w$$

Gradient for hidden layer weight:

$$\frac{\partial L}{\partial w_{1,1}^{(h)}} = \frac{\partial L}{\partial a_1^{(out)}} \cdot \frac{\partial a_1^{(out)}}{\partial a_1^{(h)}} \cdot \frac{\partial a_1^{(h)}}{\partial w_{1,1}^{(h)}} + \frac{\partial L}{\partial a_2^{(out)}} \cdot \frac{\partial a_2^{(out)}}{\partial a_1^{(h)}} \cdot \frac{\partial a_1^{(h)}}{\partial w_{1,1}^{(h)}}$$

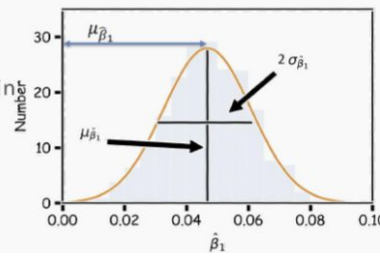
Standardized coeffs

$$t_j^{\text{boot}} = \frac{b_j}{SD^{\text{boot}}(b_j)} \quad t_j = \frac{b_j}{SD(b_j)}$$

To do so, we define a new metric, which we call t-test statistic:

$$t = \frac{\mu_{\hat{\beta}_1}}{\sigma_{\hat{\beta}_1}}$$

which measures the distance from zero in units of standard deviation.



Norm Penalties

We used to optimize:

$$W^{(i+1)} = W^{(i)} - \lambda \frac{\partial L}{\partial W}$$

Change to:

$$L_R(W; X, y) = L(W; X, y) + \frac{1}{2}\alpha \parallel W \parallel_2^2$$

$$W^{(i+1)} = W^{(i)} - \lambda \frac{\partial L}{\partial W} - \boxed{\lambda \alpha W^{(i)}}$$

Weights decay in proportion to size

Biases not penalized

L_2 regularization:
- Weights decay

$$\Omega(W) = \frac{1}{2} \parallel W \parallel_2^2$$

L_1 regularization:
- encourages sparsity

$$\Omega(W) = \frac{1}{2} \parallel W \parallel_1$$

To compare the t-test values of the predictors from our model, $|t^*|$, with the t-tests, calculated using random data, $|t^R|$, we estimate the probability of observing $|t^R| \geq |t^*|$.

We call this probability the p-value.

$$p - value = P(|t^R| \geq |t^*|)$$

In a formula, letting v being the payout or value function, the Shapley Value for the feature i can be computed as:

$$\phi_i(v) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!}[v(S \cup i) - v(S)]$$

where v(S) is the prediction for feature values in set S that are marginalized over features that are not included in set S minus the average prediction:

1. State Hypothesis:

Null hypothesis: There is no relation between X and Y

The alternative: There is some relation between X and Y

2. Choose test statistics

t-test

3. Do permutation testing

4. **Reject or not reject the hypothesis:**

We compute ***p-value*** **,** the probability of observing any value equal to $|t|$ or larger, from random data.

p-value < p-value-threshold **we** reject the null.

Bw Algo Posterior Decoding

$$P(\pi_i = k \mid x) = \frac{f_k(i) \cdot b_k(i)}{P(x)}$$

choose?

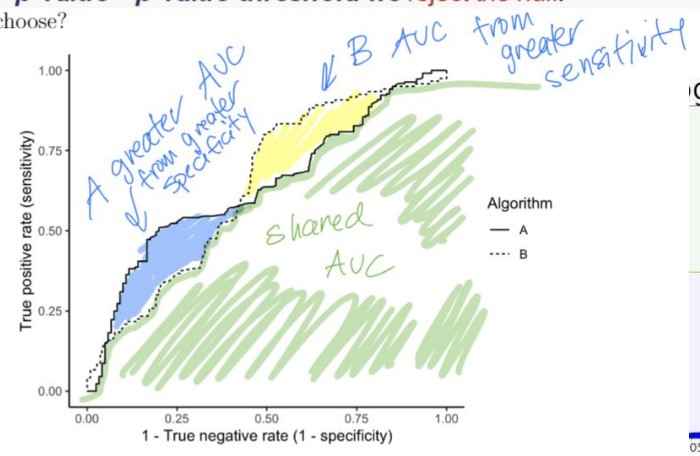
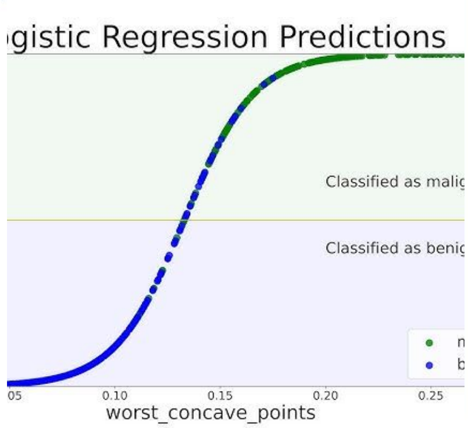


Figure 3: Algorithm ROC Curves



Basic Derivatives Rules

Constant Rule: $\frac{d}{dx}(c) = 0$

Constant Multiple Rule: $\frac{d}{dx}[cf(x)] = cf'(x)$

Power Rule: $\frac{d}{dx}(x^n) = nx^{n-1}$

Sum Rule: $\frac{d}{dx}[f(x) + g(x)] = f'(x) + g'(x)$

Difference Rule: $\frac{d}{dx}[f(x) - g(x)] = f'(x) - g'(x)$

Product Rule: $\frac{d}{dx}[f(x)g(x)] = f(x)g'(x) + g(x)f'(x)$

Quotient Rule: $\frac{d}{dx}\left[\frac{f(x)}{g(x)}\right] = \frac{g(x)f'(x) - f(x)g'(x)}{[g(x)]^2}$

Chain Rule: $\frac{d}{dx}f(g(x)) = f'(g(x))g'(x)$