A Page To Pass Machine Learning

Information Gain

$$Entropy(S) = -\sum_{i=1}^{c} p_i \log p_i$$

$$IG(S, A) = Entropy(S) - \sum_i \frac{|S_{v}|}{|S|} Entropy(S_v)$$

$$Gini(A) = \sum_i p_i \cdot (1 - p_i) \text{ or } \sum' p_i \sum p_j$$

CART uses gini, C4.5 uses IG

CART is always binary, handles missing vals nicely

With ID3 decision trees, always take the attribute with the highest information gain, as it reduces entropy the most

- not optimal as greedy
- can overfit, so choose small trees
- harder on continuous data

Scoring

$$Accuracy = \frac{Correct}{All}$$

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

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

$$F1\ Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

$$True\ Positive\ Rate = \frac{TP}{TP+FN}$$

$$True\ Negative\ Rate = \frac{TN}{TN+FP}$$

ROC curve plots true positive against true negative, AUC is area under curve, 1 is best, 0.5 is worst

Logistic Regression

$$\log \frac{p(x)}{1-p(x)} = \sum b_j x_j$$

$$\frac{p(x)}{1-p(x)} = \exp(\sum b_i x_i)$$

$$p(z) = \frac{\exp z}{1+\exp z}, \quad z = \sum b_i x_j$$

$$= (\exp z)(1+\exp z)^{-1}$$

$$p'(z) = (\exp z)(1+\exp z)^{-1} + (\exp z)(-1)(1+\exp z)^{-2}(\exp z)$$

$$= \frac{\exp z}{1+\exp z} - \frac{(\exp z)^2}{(1+\exp z)^2}$$

$$= \frac{(\exp z)(1+\exp z)}{(1+\exp z)^2} - \frac{(\exp z)^2}{(1+\exp z)^2}$$

$$= \frac{\exp z}{(1+\exp z)^2}$$

$$= \frac{\exp z}{1+\exp z} \cdot \frac{1}{1+\exp z}$$

$$= p(z)\,(1-p(z))$$

$$\log \frac{p(x)}{1-p(x)} = \sum b_j x_i$$

$$\frac{p(x)}{1-p(x)} = \exp \sum b_j x_j$$

$$p(z) = \frac{\exp z}{1+\exp z}, \quad z = \sum b_i x_j$$

$$= \frac{\exp z}{1+\exp z} \cdot \frac{\exp -z}{\exp -z}$$

$$= \frac{1}{1+\exp -z}$$

likelihood $\prod p(z) \prod (1-p(z))$

Bias-Variance

Bias error from model assumptions (underfit)

Variance fluctuation/deviation from mean (overfit)

Tradeoff means reducing one will increase another (eg. high bias, low variance)

Generative and Discriminative

Generative models distribution of actual data

e.g. naive bayes $p(t|X, w, \sigma^2) = \prod N(x_n w, \sigma^2)$

Discriminative calculates decision boundaries of classes

e.g. logistic regression $\prod p(z) \prod (1-p(z))$

- D slower processing, may consider all data
- D often has higher accuracy
- G less likely to overfit on small datasets

Independently and Identically Distributed

Same probability distribution

all mutually independent

Distance Metrics

Manhattan distance $= \sum |x_i - y_i|$

Euclidean distance $= \sqrt{\sum (x_i - y_i)^2}$

Hamming distance $= \sum \left\lceil \frac{x_i + y_i}{2} \right\rceil$

Least Squares

$$L = \frac{1}{N} \sum (t_n - \omega^T x_n)^2$$
$$= \frac{1}{N} (t - X\omega)^T (t - X\omega)$$
$$= \frac{1}{N} (X\omega - t)^T (X\omega - t)$$
$$= ((X\omega)^T - t)^T (X\omega - t)$$
$$= \frac{1}{N} (X\omega)^T X\omega - \frac{1}{N} (X\omega)^T t - \frac{1}{N} t^T X\omega + \frac{1}{N} t^T t$$
$$= \frac{1}{N} \omega^T X^T X \omega - \frac{2}{N} \omega^T X^T t + \frac{1}{N} t^T t$$
$$= \frac{1}{N} (\omega^T X^T X \omega - 2\omega^T X^T t + t^T t)$$
$$\frac{dL}{d\omega} = \frac{2}{N} X^T X \omega - \frac{2}{N} X^T t = 0$$
$$X^T X \omega = X^T t$$
$$I\omega = (X^T X)^{-1} X^T t$$
$$\hat{\omega} = (X^T X)^{-1} X^T t$$

K-Means

initialise K random centroids

for each iteration:

  for each data point:

    Calc distance to all centroids

    assign to cluster of nearest centroid

  Calc new centroid as mean of points

Reduce Overfitting in CNN

- Use more data
- Add regularisation
- Reduce number of parameters
- Reduce connections among fully connected layers

Limitations of PCA

- Assumes data is real, continuous, and no missing values
- Assumes variance shows what is interesting in data
- Assumes data is Gaussian distributed

Principles Components Analysis

Reduces the dimensionality of data

Principle components are underlying structure of data, found by finding directions of most variance

Want to find maximum eigenvalue, its corresponding eigenvector is the principle component

Eigenvalues amount equal to original dimensionality, but then remove small eigenvalues