COMP9417 – Week 5 Tutorial notes

Non-parametric Methods

Parametric modelling

We make assumptions on the type of function which our data takes

- Linear regression
- Perceptron
- Logistic regression

Non-parametric modelling

We make no assumptions on the underlying function and purely use our datapoints as guides for pattern inference

- k-Nearest neighbours
- Local regression
- Decision Trees

Entropy

Entropy measures the uncertainty of a random variable

$$H(S) = \sum_{x \in X} -p(x) \log p(x)$$

where $p(x)$ represents the proportion of x in S

If $X \sim \text{Bernoulli}(p)$:

$$H(X) = -(1-p)\log(1-p) - p\log p$$

Gain

To measure the information we gain by splitting on an attribute A for a dataset S, we define:

$$\text{Gain}(S,A) = \text{Current entropy} - \text{Entropy if we split on } A$$

If we have a dataset S with a feature A:

$$\text{Gain}(S,A) = H(S) - \sum_{v \in V_A} \frac{|S_v|}{|S|} H(S_v)$$

Basic Example

Say we have a dataset as follows $[29+, 35-]$

- $A1 \sim T: [21+, 5-]$ $F: [8+, 30-]$
- $A2 \sim T: [18+, 33-]$ $F: [11+, 2-]$

$$H(S) = \sum_{x \in X} -p(x) \log p(x)$$

$$= -\frac{29}{29+35} \log\left(\frac{29}{29+35}\right) - \frac{35}{29+35} \log\left(\frac{35}{29+35}\right)$$

$$= 0.9936$$

Dataset: $[29+, 35-]$

$H(S) = 0.9936$

- $A1 \sim T: [21+, 5-]$ $F: [8+, 30-]$

$$H(S_{A1,T}) = -\frac{21}{26} \log\left(\frac{21}{26}\right) - \frac{5}{26} \log\left(\frac{5}{26}\right)$$

$$= 0.7063$$

$$H(S_{A1,F}) = -\frac{8}{38} \log\left(\frac{8}{38}\right) - \frac{30}{38} \log\left(\frac{30}{38}\right)$$

$= 0.7425$

Dataset: [29+, 35-]:

- A2 ~ T: [18+, 33-] F: [11+, 2-]

$H(s) = 0.9936$
$H(S_{A1,T}) = 0.7063$
$H(S_{A1,F}) = 0.7425$

$$H(S_{A2,T}) = -\frac{18}{51}\log\left(\frac{18}{51}\right) - \frac{33}{51}\log\left(\frac{33}{51}\right)$$

$$= 0.9366$$

$$H(S_{A2,F}) = -\frac{11}{13}\log\left(\frac{11}{13}\right) - \frac{2}{3}\log\left(\frac{2}{13}\right)$$

$$= 0.4674$$

Dataset: [29+, 35-]

- A1 ~ T: [21+, 5-] F: [8+, 30-]
- A2 ~ T: [18+, 33-] F: [11+, 2-]

$H(s) = 0.9936$
$H(S_{A1,T}) = 0.7063$
$H(S_{A1,F}) = 0.7425$
$H(S_{A2,F}) = 0.9366$
$H(S_{A2,F}) = 0.4674$

$$\text{Gain}(S, A_1) = H(s) - \sum_{v \in \{T,F\}} \frac{|A_{1,v}|}{|S|} H(A_{1,v})$$

$$= H(S) - \frac{26}{64} H(A_{1,T}) - \frac{38}{64} H(A_{1,F})$$

$$= 0.2658$$

$$\text{Gain}(S, A_2) = H(s) - \frac{51}{64} H(A_{2,T}) - \frac{13}{64} H(A_{2,F})$$

$$= 0.1643$$

<u>ID3 Algorithm</u>

Basically what we just did:

- Calculate the entropy for each attribute $a \in A$
- Split on the attribute with the maximum Gain. This means creating a decision tree node using that attribute
- Recurse this new subset of the data.

k-NN

We predict $\hat{y}_i$ for a point x_i to be the average of the k-nearest points

Regression If we define the set k as the k-nearest neighbors of a point X_i, then our k-NN estimate is:

$$\hat{y}_i = \frac{1}{k}\sum_{i=1}^{n} 1\{x_i \in k\} y_i$$

Classification we assign X_i the majority class in k.

Linear Smoothing

k-NN regression typically fits a choppy model to our data. Linear smoothing tries to smooth out the fit by incorporating a kernel to weight the influence nearest neighbours by distance.

If we define h as the smoothing parameter and K as the kernel, the Linear Smoothing estimate is:

$$\hat{y}_i = \frac{\sum_{i=1}^{n} k\left(\frac{\|x_i - x_j\|}{h}\right) y_i}{\sum_{j=1}^{n} k\left(\frac{\|x_i - x_j\|}{h}\right)}$$

As $h \to 0$ our distances have a higher variance. If $h \to \infty$ have a lower variance, and our model is in turn smoother