(17th March, Week 3) *note we continue with some basics around different scenarios from last lecture, worth reviewing come final preparation time*

Regression

Regression is effectively an extension to our understanding of decision tree classifiers, where the final values presented are numerical rather than label like. Best described via a single feature dimension say $x_i = y_i$, where for each i we iterate through the sample space and let y_i be our labelled *numerical* value.

Discretize

The method of approach here is to effectively bucket the distribution painted by the collection of sample labels and just apply techniques that we already have learnt in previous lectures to build decisions based on entropy and the such. Ofcourse we can say that;

- Course discretization reduces overall resolution of the resolved regression "curve", whereas
- Fine discretization requires plenty of samples (to ideally have coverage on each bucket).

Linear Regression

Linear regression makes predictions using the fabled $wx_i = \hat{y}_i$, the parameter w representing a "weight" to x_i. We forgo the need to include a heigh displacement ("y-intercept") for now...

Least Squares

Ofcourse the key fitness function used is the "least square sums" function

$$\sum_{i=1}^{n}(wx_i - y_i)^2$$

Let it be *some* function $f(w)$.. to minimize this sum, we need to iterate through w and find the least $f(w)$. This "coincidentally" solves the key issue of finding the best "curve fit" w we can use for extrapolated prediction making.

$f'(w) = 0$

Assuming that of course for all i that x_i and y_i is defined, we are simply trying to find a w such that the rate of change of this error sum "flips", from an increasing sum to a decreasing sum. This emphasises a minima (and maxima

hence important to decipher which candidate w) hence yielding your candidate w to use.

$$\begin{aligned} f(w) = \frac{1}{2}\sum_{i=1}^{n}(wx_i - y_i)^2 &= \frac{1}{2}\sum_{i=1}^{n}(w^2x_i^2 - 2wx_iy_i + y_i^2) \\ &= \frac{w^2}{2}\sum_{i=1}^{n}x_i^2 - w\sum_{i=1}^{n}x_iy_i + \frac{1}{2}\sum_{i=1}^{n}y_i^2 \\ &= \frac{w^2}{2}a - wb + c \end{aligned}$$

Ofcourse a, b, c just being new variable instantiations for structural simplicity in the following deduction which is;

$$\begin{aligned} f'(w) &= wa - b \\ \Rightarrow_{f'(w)=0} w &= \frac{b}{a} = \frac{\sum_{i=1}^{n}x_iy_i}{\sum_{i=1}^{n}x_i^2} \end{aligned}$$

Also note that $f''(w)$ is strictly positive which in turn suggests what we have above is the minimal w.

Generalisation

So indeed we just looked at single variate linear regression, but in reality wed́ like to associate various "features" to a numerical label. Hence we can begin to apply the same strategy as above to multivariate predictions such as $\hat{y}_i = w_1x_{i_1} + w_2x_{i_2}$ and so on...

Vectorisation

And so it comes to no surprise that for each w_i we can form the vector w^T which is simply a compressed d-dimensional vector that summarise the weights of each $x_{i_l}, ..., x_{i_d}$ on $\hat{y}_i$

$$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

Hence $\hat{y}_i = w^Tx_i$

Thus the linear least squares model in d-dimensions minimizes the familiar

$$f(w) = \frac{1}{2}\sum_{i=1}^{n}(wx_i - y_i)^2$$

We must deduce the best w by evaluating the partial derivates wrt. each $w_1, ..., w_d$. Respectively set to zero.