Maximum Likelihood and Classification

Maximum Likelihood:

Maximum Likelihood fundamentally uses a probability distribution over a dataset.

$$P(x_i|\theta) \qquad P(x_i, ..., x_N|\theta) = \prod_{i=1}^{N}(P(x_i|\theta)) \qquad L(\theta) = \sum_{i=1}^{N} log(P(x_i|\theta))$$

The idea is to change theta in order to increase the value outputted by the probability distribution. The function is all wrapped in a log since it's easier to do calculus with. The maximum log-likelihood = maximum likelihood.

Example - Biased coin toss:

1 = heads, 0 = tails

$$P(x = 1|p) = p \qquad P(x = 0|p) = 1 - p$$

$$P(x|p) = p^x(1 - p)^{1-x}$$

$$L(p) = \sum_{i=1}^{N} log(p^{x_i}(1 - p)^{1-x_i})$$

$$L(p) = \sum_{i=1}^{N}(x_i log(p) + (1 - x_i)log(1 - p))$$

$$L(p) = (\sum_{i=1}^{N} x_i)log(p) + (N - \sum_{i=1}^{N}(x_i))log(1 - p)$$

From this function, we can differentiate with respect to L(p) to find the value of p that gives the highest likelihood value.

$$0 = \frac{\delta L(p)}{\delta p} \qquad \frac{\delta log(p)}{\delta p} = \frac{1}{p} \qquad \frac{\delta log(p)}{\delta p} = -\frac{1}{1 - p}$$

$$0 = (\sum_{i=1}^{N} x_i)\frac{1}{p} - (N - \sum_{i=1}^{N} x_i)\frac{1}{1 - p}$$

$$0 = (\sum_{i=1}^{N} x_i)(1-p) - (N - \sum_{i=1}^{N} x_i)p$$

$$0 = (\sum_{i=1}^{N} x_i) - Np \qquad Np = \sum_{i=1}^{N} x_i \qquad p = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Classification:

The data is in the form of N input vectors that contain scalar values representing its feature points and N binary targets which label each scalar into either class.

L(w) = Number of wrong classifications (loss function)

$$f_w(x_i) = \theta(w^T x_i) = \theta(\sum_{j=1}^{n} w_j X_{ij})$$

$$\theta(a < 0) = 0 \qquad \theta(a \geq 0) = 1$$

Since this loss function is discrete we cannot use it to perform calculus to find the minimum point. We can use a probability function instead.

$$P = (y_i = 1|x_i) = \sigma(w^T x) = \sigma(\sum_{j=1}^{n} X_{ij} w_j)$$

Where sigma represents a continuous sigmoid function:

$$\sigma(a) = \frac{1}{1+e^{-a}}$$

$y_i = 1$ is more likely if $w^T x > 0$

$y_i = 0$ is more likely if $w^T x < 0$

Gaussian:

$$p(x|m,v) = \frac{1}{\sqrt{2\pi v}} e^{-\frac{(x-m)^2}{2v}}$$

$$log(p(x|m,v) = log(\frac{1}{2\pi}) + log(\frac{1}{\sqrt{v}}) - \frac{(x-m)^2}{2v}$$

$$= -\frac{1}{2}log(2\pi) - \frac{1}{2}log(v) - \frac{1}{2v}(x-m)^2$$

$$log(p(x_1,...,x_n) = \sum_{i=1}^{N}(log(p(x_i|m,v)))$$

$$= \sum_{i=1}^{N}(log(\frac{1}{2\pi}) + log(\frac{1}{\sqrt{v}}) - \frac{(x_i-m)^2}{2v})$$

$$= -\frac{N}{2}log(2\pi) - \frac{N}{2}log(v) - \frac{1}{2v}\sum_{i=1}^{N}(x_i-m)^2$$

$$\frac{\delta log(P(x_1,...,x_N|m,v)}{\delta m} = 0 \qquad \frac{\delta log(P(x_1,...,x_N|m,v)}{\delta v} = 0$$

$$0 = \frac{\delta log(p(...))}{\delta m} = -\frac{1}{2v}\frac{\delta}{\delta m}\sum_{i=1}^{N}(x_i-m)^2 = \frac{\delta}{\delta m}\sum_{i=1}^{N}(x_i-m)^2$$

The above function is differentiating the squared distance between X and some parameter m which is a constant representing the mean, therefore this problem is the same as "yconst".

$$m = \frac{1}{N}\sum_{i=1}^{N}x_i = Mean$$

$$0 = \frac{\delta log(p(...))}{\delta v} = -\frac{N}{2v} + \frac{1}{2v^2}\sum_{i=1}^{N}(x_i-m)^2$$

$$\frac{Nv}{2} = \frac{1}{2}\sum_{i=1}^{N}(x_i-m)^2 \qquad \frac{1}{N}\sum_{i=1}^{N}(x_i-m)^2 = Variance$$