|

Assignment - 3

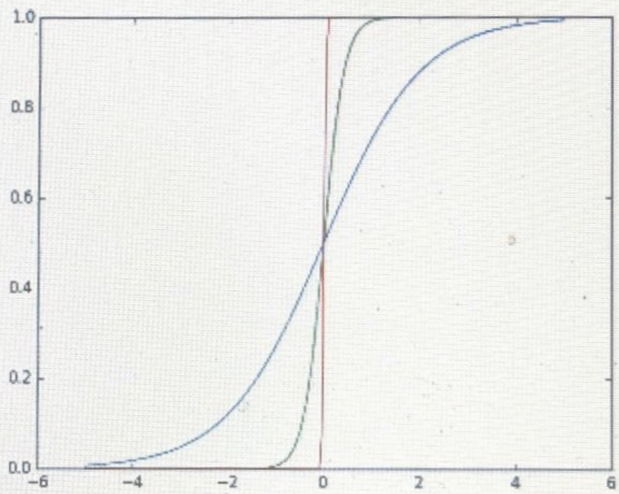
AML

Saurav vara prasad Channuri
ES16BTECH11007

Q1)

a.

As the value of the regularization parameter is increasing the curve is getting straighter and the slope is going on increasing as the weights are going on increasing



With large weights even a little change in the input values will lead to large changes in the probabilities. Hence the curve overfits. Even the input values with very little difference can be classified into two different classes.

1A) ~~3A~~) Given, the prior on weights is $P(\omega_0, \_\_, \omega_d)$

b) MAP estimate is written as

$$W_{MAP} = \max_{\omega_0, \_, \omega_d} \prod_{i=1}^{n} P(y_i/x_i, \omega_0, \_\_, \omega_d) P(\omega_0, \_, \omega_d)$$

Now, assuming the prior to be following gaussian distribution

$$\log(W_{posterior}) = \log\left(P(\omega) \prod_{j=1}^{n} P(y^j/x^i, \omega)\right)$$

$$P(\omega) = \prod_{i=0}^{d} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\omega_i^2}{2}\right)$$

$$W_{MAP} = \arg\max_{\omega} \log(W_{posterior}) = \arg\max\left[\sum_{j=1}^{n} \log(P(y^j/x^i, \omega)) - \sum_i \frac{\omega_i^2}{2}\right]$$

the gradient ~~descent~~ ascent can be updated as

$$\omega_i^{(t+1)} = \omega_i^{(t)} + \alpha \left.\frac{\partial L(\omega)}{\partial \omega_i}\right|_t$$

where

gradient of the log of posterior prob. distribution is given by

$$\frac{\partial L(\omega)}{\partial \omega_i} = \underbrace{\frac{\partial}{\partial \omega_i}\left(\log(P(\omega))\right)} + \frac{\partial}{\partial \omega_i} \log\left(\prod_{j=1}^{n} P(y^j/x^i, \omega)\right)$$

$$\Rightarrow \frac{\partial}{\partial \omega_i}\left(\log(P(\omega))\right) = -\omega_i$$

the final update rule can be written as

$$\omega_i^{(t+1)} = \omega_i^{(t)} + \eta\left[-\omega_i^{(t)} + \sum_j x_i^j \left(y^j - P(Y=1/x^j, \omega^{(t)})\right)\right]$$

c) we know that all the probabilities of classes sum to 1

$$P(Y=y_K/x) = 1 - \sum_{k=1}^{K-1} P(Y=y_k/x)$$

and $P(Y=y_k/x) \propto \exp\left(\omega_{k0} + \sum_{i=1}^{d} \omega_{ki} X_i\right)$ $\quad k = 1, \_, K-1$

Since, introducing another set of weights is redundant

we can define $$P(Y=y_k/x) = \frac{\exp(\omega_{k0} + \sum_{i=1}^{d} \omega_{ki} x_i)}{1 + \sum_{k=1}^{K-1} \exp(\omega_{k0} + \sum_{i=1}^{d} \omega_{ki} x_i)}$$ → just as in Binary classification and the classification rule can simply pickup the label with highest probability

$$y = y_{k^*} \quad \text{where } k^* = \underset{k \in \{1, .., K\}}{\arg\max} \; P(Y = y_k / x)$$

d) The decicion boundary between each pair of classes is linear and hence the overall decision boundary is piecewise linear

Since $\arg\max_i \exp(a_i) = \arg\max_i a_i$ and max of linear functions is piecewise-linear, the overall decision boundary is piecewise linear

2A) a) $k(x_i, x) = \exp\left(\frac{-\|X_i - X\|^2}{2\sigma^2}\right)$

$$\hat{y} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \quad \text{where } w_i = \exp\left(-\frac{D(x_i, x)^2}{k_w^2}\right)$$

let $l_i(x) = \frac{w_i}{\sum_{i=1}^{n} w_i}$. Then $\hat{y} = l(x)^T y$

hence the kernel regression is a linear smoother

b) No, It is not a linear smoother

In general, there is no closed form solution for w that minimizes sum of absolute values of the errors. Yet, the solution can be seen to be similar to a median

An optimal 'w' makes the same number of positive errors as negative errors.

Counter example is constant input where each training point has $x_i = 1$ for a variety of 'y' values, so, 'w' is median of y's.

Clearly w is not linear in any of the y's, hence the median changes as rank of y's does

c) for $x \in B_i$, $l_j(x) = \frac{I(x_j \in B_i)}{|B_i|}$, hence the regressogram is a linear smoother