CSCI 4360/6360 Data Science II
Department of Computer Science
University of Georgia

Assignment 2: Guest Lecturer Edition

Aditya Shinde

September 14, 2017

Questions

Linear Regression **[20pts]**

For this question, I watched a couple of videos on youtube about maximum liklihood estimation and sampling data points from probability distributions.

1

If the error terms ϵ_i are assumed to be i.i.d and sampled from a Gaussian distribution,

$$\epsilon \sim N(0, \sigma^2)$$
$$\epsilon = y - w^T x$$
$$y \sim N(w^T x, \sigma^2)$$

Similarly, If

$$\epsilon \sim Lap(0, b)$$
$$\epsilon = y - w^T x$$
$$y \sim Lap(w^T x, b)$$

$$P(Y|x_i;w) = \text{argmax}\Pi_i Lap(w^T x, b)$$

$$P(Y|x_i;w) = \text{argmax}\Pi_i \frac{1}{2b} e^{(\frac{-|y-w^T x|}{b})}$$

$$\log P(Y|x_i;w) = \text{argmax} \sum -\log 2b + \left(\frac{-|y-w^T x|}{b}\right)$$

Now, $\log 2b$ is a constant,
So we can write $J_{Lap}(w)$ as,

$$J_{Lap}(w) = \text{argmin} \sum \left(|y-w^T x|\right)$$

2

If the noise terms are distributed as Gaussians, the loss function is in the form of squared errors. This places a very high penalty on outliers and thus the model is highly reactive towards them. In case of a Laplacian distribution, the loss function is the mean absolute error. This is less reactive to outliers.

Regularization **[30pts]**

1

As λ decreases, the cost function penalises exploding weights to a lesser extent.
As $\lambda \to 0$ the cost function $J_R(\beta) \to J(\beta)$
As λ increases, the regularization term becomes more dominant. It starts heavily penalising all weights. As $\lambda \to \text{inf}$, the cost function $J_R(\beta) \sim \lambda||\beta||^2$
The model will only fit the intercept.

2

$$\beta_{\text{MAP}} = \text{argmax}_\beta \prod_{i=1}^{n} P(Y_i|X_i;\beta)P(\beta)$$

$$\log \beta_{\text{MAP}} = \text{argmax}_\beta \sum_{i=1}^{n} (\log P(Y_i|X_i;\beta) + \log P(\beta))$$

$$\log \beta_{\text{MAP}} = \text{argmax}_\beta \sum_{i=1}^{n} \left(-(Y-W^T X)^2 + (-\log \frac{\lambda}{I\sigma^2} - \frac{\beta^2 \lambda}{I\sigma^2})\right)$$

Getting rid of the constants.

$$\text{argmax}_\beta \log \beta_{\text{MAP}} = \text{argmax}_\beta \sum_{i=1}^{n} - \left((Y-W^T X)^2 + \frac{\beta^2 \lambda}{I\sigma^2}\right)$$

If we assume unit variance, $\sigma^2 = 1$ and I is an identity matrix. And multiply and divide right hand side by -1 to change argmax to argmin.

$$\text{argmax}_\beta \log \beta_{\text{MAP}} = \text{argmin}_\beta \sum_{i=1}^{n} \left((Y - W^T X)^2 + \lambda \beta^2\right)$$

3

Now in this case β is a random variable with variance $\frac{I\sigma^2}{\lambda}$.
If $\lambda \to 0$ in this case, that means $Var[\beta] \to \text{inf}$, that is variance tends to infinity. So the samples will be far away.
Similarly, if $\lambda \to \text{inf}$, $Var[\beta] \to 0$ which means $p(\beta)$ will be a single point at 0 and we would essentially be fitting only the intercept.

Evolutionary computation

1

Generational genetic algorithms use specific distint generations as different sets of parameters. By keeping distinct generations across epochs, we make sure that none of the individuals from the previous generation are in the current generation. This is specially helpful when the cost function is dynamic. It can also be helpful to find multiple solutions.
Steady state GAs on the other hand use a replacement mechanism. The newer individuals which are generated using genetic operators are swapped with the weaker individuals of the current generation.
One of the advantages of steady state GA's is that they require less resources in terms of memory since the population size is constant. Generational GA's on the other hand need twice as much memory. But steady state GA's usually take more time to converge.

2

Pittsburgh and Michigan classifers both work on the same principles of evolutionary computation. The distinction is in the role of a single individual. In the Pittsburgh classifiers, each individual is a complete solution consisting of many rules. In case Michigan classifiers, an individual is just an element of a rule based system. Many such individuals need to be combined to make a complete solution.
One of the advantages that Michigan classifers have is diversity. Since the individuals are made of elementary rules, they can be used with many problems easily as compared to Pittsburgh.