Weekly report of lessons

Name: Suryam Arnav Kalra
Roll No: 19CS30050
The week: 20th September 2021 to 24th September 2021

The topics covered:

- Parametric Methods and Maximum Likelihood estimation of parameters
- Bernoulli Distribution, Multinomial Density function and Gaussian Density function
- Bias, MSE and Variance of estimators
- The Bayes' Estimator and Bayes' Estimator of mean of a normal distribution
- Parametric Classification
- Multivariate representation
- Non-parametric approaches and Univariate nonparametric density estimation
- Kernel Estimator and KNN estimator
- Nonparametric Classification and Instance Based Learning
- KNN regression and Locally weighted regression
- Class and Cluster
- K-means clustering
- The Lloyd algorithm and its Strength and Weakness
- K-means++

Summary topic wise:

- Parametric methods: The probability density functions of known form are described by a set of parameters, e.g. P(x|θ) where θ is a set of parameters .
- Maximum Likelihood estimation of parameters : Likelihood = $l(\theta|X) = P(X|\theta) = \prod P(x^t|\theta)$ and MLE of $\theta = argmax\ l(\theta|X)$.
- Bernoulli Distribution: $L(p|X) = \log \prod p^{x^t}(1-p)^{x^t}$ and after taking the derivate we have $p' = \frac{\sum x^t}{N}$
- Multinomial Density function: It is a generalization of the Bernoulli process in which one of K mutually exclusive states occurs at every trial.
- Gaussian density function: $p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2}(\frac{x-\mu}{\sigma})^2}$ with $E(X) = \mu$, $Var(X) = \sigma^2$
- Bias, MSE and Variance of estimators: Let d(X) = d be estimator of a parameter θ then Bias : $b_\theta(d) = E\big(d(X)\big) - \theta$, MSE of estimator = $E((d(X) - \theta)^2)$ and $Var(X) = E(X^2) - E(X)^2$. Also, $MSE = Variance\ of\ d + (bias\ of\ d)^2$
- The Bayes' Estimator: Estimation of θ by considering posterior probability $P(\theta|X)$
$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} = \frac{P(X|\theta)P(\theta)}{\int P(X|\theta)P(\theta)d\theta}$$
- Bayes' Estimator of mean of a normal distribution:
$x^t \sim N(\mu, \sigma^2)\ and\ \theta \sim N(\mu_0, {\sigma^2}_0)$ and Mean is given by weighted mean of μ and μ_0.
- Parametric Classification: First, compute the posterior for all classes and then assign the class with the maximum posterior.

- Multivariate representation: It is in the form of a data matrix where each row is a data sample. We have mean vector $\boldsymbol{\mu}$ = $[\mu_1\ \mu_2\ \ldots\ \mu_d]$ and a covariance matrix $\sum = [\sigma_{ij}]$ and a correlation matrix $[\rho_{ij} = \sigma_{ij}/\sqrt{\sigma_i\sigma_j}]$
- Nonparametric approaches: The only assumption is that similar inputs have similar outputs. It estimates the probability density locally and uses instance bases learning.
- Univariate nonparametric density estimation: We first estimate the cumulative probability and the estimated probability density. We can then use a naïve estimator or a histogram of specified bin-width.
- Kernel estimator: A kernel function is a function of distance used to determine the weight of each sample, $P(x) = \frac{1}{Nh}\sum K(\frac{x - x^t}{h})$
- KNN estimator: K-nearest neighbor estimator uses the distance metric to predict the class of the given example. Let $d_i(x)$ = distance of the ith NN from x then K-NN density estimate = $k/(N(2d_k(x)))$
- Instance based learning: In this type of learning we retrieve a set of similar related instances and classify/regress using them. It has a significant advantage over complex target function, instead of computing a global function, compute locally.
- KNN regression: Let the ith neighbor of x be x_i then $f'(x) = \frac{\sum f(x_i)}{k}$. Also for weighted regression we can use weight inversely proportional to square of the distance $d(x, x_i)$.
- Locally weighted regression: The target function is linear on attribute variables and the MSE can be over unweighted KNN's , over all training examples with weight proportional to the kernel or over KNN's weight proportional to the kernel.
- Class and Cluster: Class is a well studied group of objects identified by their common properties or characteristics whereas a cluster is a group with 'loosely' defined similarity among the objects.
- K-means clustering: Compute K partitions of the data points, so that it minimized the sum of square of distances between a data point and the center of its respective partition (cluster).
- The Lloyd algorithm: Given K initial centers, assign a point to the cluster represented by its center, if it is the closest among them and then update the centers till the centers do not change their positions.
- Strength and weakness: The convergence is guaranteed at a quadratic rate and we have a linear time complexity in N, d and K. The weaknesses are: Improper initialization, may get stuck at a local minima and is sensitive to noise.
- K-means++: The first center c_1 is chosen randomly. The ith center c_i is chosen as x' with a probability proportional to square of the minimum distance from the selected i-1 centers.

Concepts challenging to comprehend:

Nonparametric approaches along with Bias and Variance of estimator are a little bit challenging to comprehend.

Interesting and exciting concepts:

KNN and K-means clustering are quite interesting and exciting to learn.

Concepts not understood:

After going through the book and the video lectures the concepts are clearly understood.

Any novel idea of yours out of the lessons:

Instead of starting with each and every point as a center in K-means clustering, we can start with all of the points as one big cluster. We can then use linear/logistic regression to help divide the cluster into two small clusters by minimizing the sum of square of distance. We can then use the same for the two small clusters till the division does not result in better accuracy. In this way, we can combine the power of two learning models as well.