Assignment 1

* Section A

1) True

2) True

3) False

4) True

5) True

6) False

7) False

8) True

9) True

10) True

* SECTION - B.

2) A: Squared loss

B: Absolute loss

C: Hinge loss

D: Logistic loss

E: 0-1 loss

3)(a) They are continous and diffruntiable, ~~allowing gradient~~.

(b) Loss function that are twice differentiable and convex, ~~such~~. can be optimised using Newton's Method because the Method requires computation of both gradients.

* SECTION C

1) Underfitting happens when a model is too simple to capture the underlying pattern in the data

2) High training and test errors indicate high ~~toss~~, bias, meaning the model is un able to capture the true relationship in the data.

3) Bagging reduces variances by averaging predictions from multiple models trained on different subsets of data.

4) Boosting reduce bias by improving model complexity ; while variance depends on noise and regularization.

* SECTION-D

1) To reduce KNN computation time
   i) Reduce data size.
   ii) Fewer dimensions
   iii) Reduce search time to O(log n).

2) Square euclidian distance does not change KNN predictions and does not solve the curse of dimensionality.

3) KNN fails in high dimensions because.
   . data ~~looks~~ become scattered, distance lose Meaning.

4) K ↑ ⇒ Bias ↑ ⇒ variance ↓
   K ↓ ⇒ Bias ↓ ⇒ variance ↑.

5) KNN is preffered over linear SVM when data has complex local structure & ~~to~~ limited size

★ SECTION - E

1) constant value → c

$$L(c) = \sum_{y=i}^{n} (y_i - c)^2$$

Differentiate w.r.t c.

$$\frac{dL}{dc} = \sum 2(c - y_i) = 0$$

$$\boxed{c = \frac{1}{n} \sum y_i}$$

2) $G = 1 - \sum_{i=1}^{k} p_i^2$

Min: $p = (1, 0, 0)$

$G = 1 - (1)^2 + 0 + 0 = 0$ → Min.

Max = $p = (1/3, 1/3, 1/3)$

$= 1 - 3\left(\frac{1}{3}\right)^2 = 2/3$ → Max.

3) Decision tress are myopic because they greedily select splits based on local impurity reduction without considering future spills.

4) ~~ovif~~ Overfitting in decision ~~trees~~ trees,

1) Pruning

2) Restrict Tree complexity.

★ SECTION-F

1) Random forest cannot use the same data for training and testing.

2) ~~Random forest should not use the same data for training and testing, but they estimate generalisation error using out-of-bag samples.~~

Bagging reduces variance

Boosting reduces bias.