Assignment 4 { Theory }

Q5

Ans

(a) Convex Sets

→ A Subset C of R^n is called convex if

$$\alpha x + (1-\alpha) y \in C$$
$$\forall x, y \in C \text{ & } 0 \leq \alpha \leq 1$$

→ operations that preserve convexity
  ↳ Intersection, scalar multiplication vector sum, closure, interior Linear transformation

A convex set is a collection of points in which the line AB connecting any 2 points A, B in the set lies completely within the set

Conven functions

Let C be a conven subset of R^n. A function $F: C \to R$ is called conven if

$$F(\alpha x + (1-\alpha) y) \leq \alpha f(x) + (1-\alpha) F(y)$$
$$\forall x, y \in C$$

If f is a conven function, then all its level sets $\{x \in C \mid F(x) \leq a\}$ & $\{x \in C \mid F(x) < a\}$, where a is a scalar are conven.

Definition : A conven function is a continuous function whose values at the midpoint of every interval in its domain does not exceed the arithmatic mean of its values at the ends of interval

A function $f(x)$ is said to be strictly conven if for every (A, B) in the domain

the line segment obtained by joining these strictly lies above the curve except the 2 end points which would be $(A, f(A))$, $(B, f(B))$ that are common

Convexity of Lasso & Ridge regression

For lasso, the lemma has been proved that, For any x, y & $\lambda \geq 0$ and lasso solutions $\beta(1)$ & $\beta(2)$ must satisfy $\beta_i^{(1)} \cdot \beta_i^{(2)} \geq 0$ for $i = 1, \cdots p$. In other words, any 2 lasso solutions must have the same signs over the common support

Hence lasso penalty is not strict because if A, B have the same sign then the line segment & curve are equal which means ∞ no. of points could be common.

Ridge regression Minimization problem is

$$J(w) = ||y - Xw||_2^2 + \lambda ||w||_2^2 \quad \text{— (1)}$$

Alternatively, it can be said that

$\hat{B} \in \text{argmin } F(XB) + \lambda ||B||$ — (2)

where the loss function $F: R^n \to R$ is differentiable & strictly conven

By (1) & (2) we can say that ridge regression is strictly conven.

Q5 (b)

Ans $K=3$ would be better than $K=2$ because we dont want to be in a situation where 1 neighbour suggest one class & another neighbour suggests another class for $K=3$, there will always be

majority because atleast 2 out of the 3 neighbours will belong to the same class, and there will be majority.

Q5 (c) {Plz see next Page}

Q5(C).

Assuming training set D consists of N points (x_i, y_i) sampled iiD

Let the classifier trained on D be H_D

$y = H_D(x)$ is a distribution with mean 'u' & and variance 'σ^2' {given}

Total Loss = $E_{xD}[L(H_D(x), y)]$

→ Loss over single x

Average loss over all 'x'

Let $L(h(x), y) = \frac{1}{2}(h(x) - y)^2$

Squared Loss.

$$E_{xD}[(H_D(x) - y)^2] = E_{xD}[(H_D(x) - E_D[H_D(x)] + E_D[H_D(x)] - y)^2]$$

$$\Rightarrow E_{x,D}[(H_D(x) - E_D(H_D(x)]^2 + (E_D(H_D(x)] - y)^2 +$$

$$2\left(H_D(x) - E_D\left(H_D(x)\right]\right)\left(E_D\left[H_D(x)\right] - y\right)\right]$$

$$\Rightarrow \boxed{\underbrace{E_{x,D}\left[\left(H_D(x) - E_D\left[H_D(x)\right]\right)^2\right]}_{\text{Variance}} + \underbrace{E_x\left[\left(E_D\left[H_D(x)\right] - y\right)^2\right]}_{\text{bias}}}$$

Suppose K models are trained on K subsets of D $\{D_i\}_{i=1}^{K}$

The bias term, given some data point (x, y), depends on $E_D[H_D(x)]$ only.

$$H_D(x) = \frac{1}{K}\sum_{i=1}^{K} H_{D_i}(x)$$

$$E_D\left(H_D(x)\right) = \frac{1}{K}\sum_{i=1}^{K} E_{D_i}\left(H_{D_i}(x)\right)$$

$$= \frac{1}{K} * K\mu$$

$$\boxed{= \mu}$$

So the bias term does not change with ensembeling.

Now $H_D(x) = \frac{1}{K} \sum_{i=1}^{K} H_{D_i}(x)$

$$\Rightarrow Var(H_D(x)) = Var\left(\frac{1}{K} \sum_{i=1}^{K} H_{D_i}(x)\right)$$

$$\left\{ Var(kx) = k^2 Var(x) \right\} = \frac{1}{K^2} Var\left(\sum_{i=1}^{K} H_{D_i}(x)\right)$$

$$\left\{ \begin{array}{l} Var(x_1 + x_2 \cdots) \\ = Var(x_1) + Var(x_2) \\ \text{when } x_i = iid \end{array} \right\} = \frac{1}{K^2} \left(\sum_{i=1}^{K} Var(H_{D_i}(x)\right)$$

$$= \frac{1}{K^2} \times KL\sigma^2$$

$$= \frac{L}{K}\sigma^2$$

Now, as the bias term does not change, we need to reduce the variance term to reduce

the loss.

thus $\frac{L}{K}\sigma^2 < \sigma^2$

$\frac{L}{K}$: variance of ensemble model.

σ^2: variance of single model trained on Entire Data

$$L < K$$

thus there must be more than L learners for the ensemble to perform better than the single bigger Model.