Q3. (b)

For naive bayes algorithm, if one of the classes has zero training samples then it will assign it zero probability and won't be able to make prediction. This is known as zero prediction problem. It skews the whole performance of the classification.

An approach to solve this problem is to add one to the count for every attribute value - class combination when an attribute value doesn't occur with every class value. This will lead to the removal of all the zero values from the classes and, at the same time, will not impact the overall relative frequency of the classes. This process of smoothing the data by adding a number is known as additive smoothing or laplace smoothing.

Qs. (i)

Convex set: A set C is convex if the line segment between any two points in C lies in C, i.e.,

$$\forall x_1, x_2 \in C$$

$$\forall \theta_1 \in [0,1]$$

$$\theta x_1 + (1-\theta) x_2 \in C.$$

Generalized definition:

A convex combination of points

$$x_1, x_2 \ldots x_k \in C$$

is any point of form

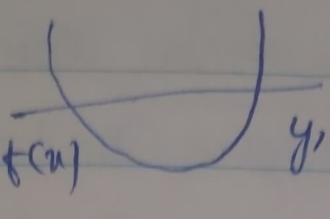
$$\theta_1 x_1 + \theta_2 x_2 + \ldots \theta_k x_k$$

where, $\theta_i \geq 0$, $i = 1, \ldots k$, $\sum_{i=1}^{k} \theta_i = 1$,

then a set C is convex if and only if any convex combination of points in C is in C.

Convex function: A convex function is a function defined on a convex domain in such that for any two points in the domain, the segment between the two points lies above the function curve between them. This can be shown in the

following figure:



graph of a convex function.

The line segment between any two points on the graph lies above the graph.

Formal definition:

A function f is convex iff its epigraph, the set of all points above the function graph is a convex set.

A function $f: R^n \to R$ is convex if dom(f) is a convex set & if $\forall x, y \in \text{dom}(f)$

$\forall \theta \in [0,1]$,

we have,

$$f(\theta x + (1-\theta) y) \leq \theta f(x) + (1-\theta) f(y)$$

The epigraph of a function $f: R^n \to R$ is the set of points

$$\text{epi}(f) = \{ (x,t) \mid x \in \text{dom}(f), \; t \geq f(x) \}$$

The ridge regression loss function is the sum of two parabolas: one is atleast convex i.e, a set $S \subset \mathbb{R}^n$ is convex if for all $\beta_1, \beta_2 \in S$ then weighted avg.

$$\beta_\theta = (1-\theta)\beta_1 + \theta\beta_2$$

$\forall\ \theta \in [0,1]$ is itself an element of S, thus $\beta_\theta \in S$.

and the other strictly convex i.e, if for all $\theta \in [0,1]$, the weighted avg β_θ is inside S & not on its boundary the set is strictly convex.

∴ The ridge regression is thus strictly convex.

The Lasso regression loss function is the sum of squares criterion and a sum of absolute functions. Both are convex in β; the former is not strictly convex due to high dimensionality and the absolute value function is convex due to its piece wise linearity.

∴ The Lasso regression is convex but not strict.

Hence, there exists multiple minimizers of the lasso loss function, they can be used to construct a convex set of minimizers.

Thus, if

$\hat{\beta}_a(\lambda_1)$ & $\hat{\beta}_b(\lambda_1)$ are lasso estimators,

then so are

$(1-\theta)\,\hat{\beta}_a(\lambda_1) + \theta\,\hat{\beta}_b(\lambda_2)$ for $\theta \in (0,1)$

Q5. (ii)

A KNN with $k=3$ would be more accurate than a KNN with $k=2$ because of the following reasons:

(a) the larger the k is, the smoother the classification boundary is.

(b) when k increases the complexity of KNN decreases.

(c) Small value of 'k' means noise will have a higher influence on the result.

(d) Odd number is always helpful in breaking the tie. This is evident from psycholinguistics and

behavioral studies that the minimum number of annotators are always 3 to get a majority voting to reduce bias & avoid garbage information.

Q5. (iii)

given: K-models on random subsets of the dataset $\{D_i\}_{i=1}^{K}$ with mean(μ) & variance of $L\sigma^2$ where $L > 1$.

Also to compute the final predictions, the predictions from the K-models are averaged.

sol,

Assuming that the models are independent, & identically distributed (iid):

$$\therefore E[y] = E\left[\frac{1}{K}\sum_{j=1}^{K} d_K\right] = \frac{1}{K} E\left[\sum_{j=1}^{K} d_K\right]$$

↙ expected value of the avg. model

$$= \frac{1}{K} K(\mu) = \mu. \quad \text{——— (1)}$$

∴ The mean μ of the k models remains same, in other words bias does not change.

Note: We know that the variance of the avg models is potentially much lower than the orijnal variance. Also bias remains the same that has been proved through equation (1).

Now, Variance:

$$Var(y) = var\left[\frac{1}{K}\sum_{i=1}^{K} d_j\right]$$

(Var(y): variance of the avg model)

$$= \frac{1}{K^2} var\left(\sum_{i=1}^{K} d_j\right)$$

$$= \frac{1}{k^2} \underbrace{var(d_1) + Var(d_2) + \cdots + Var(d_k)}_{\text{given}}$$

$$= \frac{1}{k^2} K L \sigma^2 \quad \text{(10)} \qquad = \frac{L}{K}\sigma^2 \quad \text{—(2)}$$

We know the variance of avg model should be less than variance of orijnal,

∴ we get

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

↖ Avg. model variance

↪ original variance

$$\Rightarrow \frac{L}{K} < 1$$

$$\boxed{\Rightarrow K > L} \quad \text{where } L > 1$$