Q3)b) Naive Bayes algorithm is based on Bayes theorem of Probability and statistics with an assumption that features are all independant of each other.

If one of the classes have 0 training examples then the frequency based probability estimate of Naive Bayes will be 0 and its known as the Zero Frequency Problem. We will get a 0 when probabilities are multiplied. If one feature value is not associated with a class (let class 1) in training set then in test set any ~~que~~ new query point containing that feature value will always be predicted as class 0.

Solution to above problem is to add 1 to the count of every feature/attribute value class combination so that we dont get a 0 anymore and at the same time not impact the overall relative frequency of the classes. The process ~~is~~ of ~~coined as smoothnig~~ smoothning by adding a number is called additive smoothning or Laplace smoothning.

Naive bayes will be helpful because —

✓ it works very fast

✓ best suitable for multi class problem

✓ if assumption of independance of features holds true it will perform better than other models with less training data

✓ best suited for categorical input variables

✓ used in realtime application as it is very fast

✓ used in sentiment analysis

Q5) a) Convex Set

it is a subset that intersects every line into a single line segment (possibly empty).

given ~~#~~ n points $v_1, \cdots v_n$ in a convex set S, and k nonnegative numbers $\lambda_1, \cdots \lambda_k$ such that $\lambda_1 + \lambda_2 + \cdots \lambda_k = 1$, the affine combination $\sum_{k=1}^{r} \lambda_k \mu_k$ belongs to S. As the definition of a convex set is the case $r = 2$, this ~~share~~ property characterizes the convex sets. Such an affine combination is called a convex combination of $v_1, \cdots v_n$.

conver function

a convex function is a continuous function whose value at the mid point of every interval in its domain does not exceed the arithmatic mean of its values at the ends of the interval.

mathematically,
for a convex function f(x) with interval [a, b], x_1, x_2 any 2 points in [a, b] and λ (any) $0<\lambda<1$

$$f[\lambda x_1 + (1-\lambda) x_2] \le \lambda f(x_1) + (1-\lambda) f(x_2)$$

f(x) will be strictly convex if the above inequality ~~holds for~~ is strict for all x_1, x_2.

if f(x) has a 2nd derivative then a necessary and sufficient condition for it to be convex on that particular interval is the 2nd derivative must be greater than or equal to 0 for all x in [a, b].

ridge ~~regression~~ regularization

ridge ~~regression~~ regularization is strictly convex.

defining ridge regression —

$$\min_\beta \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 , \lambda \ge 0$$

On differentiating above function twice with β we get

$$2\lambda I + 2X^T X \quad — ①$$

now ✓ λI is positive definite for any $\lambda > 0$

✓ $X^T X$ is positive definite for any X

✓ sum of a positive definite and positive definite matrix is positive definite.

Thus ① is positive definite for $\lambda > 0$

Thus we have proved the 2nd derivative of a function is strictly greater than 0 & hence the function is strictly convex.

lasso regularization

Lasso regularization is not strictly convex but only a convex function.

It is so because the penalty in lasso is not strict as here x_1 and x_2 (defined above in convex function) are of same sign and thus the line segment and the curve are exactly equal and so there are infinitely many points in common.

05) b) k=3 will be a more accurate nearest neighbour as with k=2 it might lead to overfitting. Though ~~it is problem depend~~ the two values of k are quite close and the difference it is making with respect to the cluster formation is problem dependent, still k=3 (out of the two k's) would be a better choice ; looking at only 2 neighbours might not be a good measure to classify a data point to one particular cluster, ~~we~~ we might be biased.

Also choosing k as an odd ~~is to~~ number is beneficial as if suppose we have binary classificati ~~then~~ and with k=2 we get equal votes for positive and negative class then any one class will be chosen randomly. but with k=3 this condition will never occur, we will always have a class with more votes than the other. (here votes are the number of nearest neighbours belonging to a particular class)

Q5) c) Assuming our training set is D and it consist of N points (x_i, y_i) and it is sampled IID.

Let the classifier be denoted as H_T

$y = H_D(x)$ with mean μ & variance σ^2 as given.

now we create k models such that each works on a subset of D. each has a output with mean μ and variance $k\sigma^2$.

Average total loss = $E_{xD}[L(H_D(x), y)]$

squared loss -

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

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

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

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

$\Rightarrow$ variance - bias

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

$$\Rightarrow E(H_D(x)) = \frac{1}{k}\sum_{i=1}^{k} E_{D_i}(H_{D_i}(x))$$

$$= \frac{1}{k} \times k\mu$$

$$= \mu$$

so with ensembling, bias does not change

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

$$\Rightarrow \mathrm{Var}(H_D(x)) = \mathrm{Var}\left(\frac{1}{k} \sum_{i=1}^{k} H_D(x)\right)$$

$$\Rightarrow \frac{1}{k^2} \mathrm{Var}\left(\sum_{i=1}^{k} H_{D_i}(x)\right) \quad \left\{ \mathrm{Var}(kn) = k^2 \mathrm{Var}(n) \right.$$

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

{ wen n_i is IID

$\mathrm{Var}(n_1 + n_2 \ldots) = \mathrm{Var}(n_1) + \mathrm{Var}(n_2)$

$$\Rightarrow \frac{1}{k^2} \times k L \sigma^2$$

$$\Rightarrow \frac{L}{k} \sigma^2$$

as bias is not changing, so reducing variance as our ensemble model should be probabilistically better than our original model.

$$\therefore \frac{L}{k} \sigma^2 < \sigma^2$$

$$\Rightarrow L < k$$

so range of k is $\underline{k > L}$.