CSP - 571 Final

(1) In order to have 3-cutpoints for piecewise polynomial for regression we have following constraints

① The fitted curve must be continuous

② First derivative of piecewise polynomial are continuous

③ Second derivative of piecewise polynomial are continuous

The degree of freedom for cubic spline = K+4

$\Rightarrow 3+4 = 7$

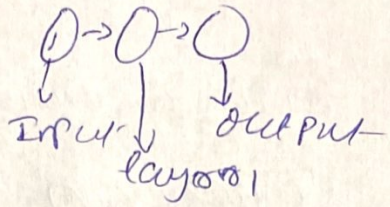
(2) Here were given d=250, N=5,000,000

we will get covariance matrix = 250×250

Implies 250 eigen values & 250 eigen vectors

→ If we chose top 10% of eigen vectors we will get data sample matrix of 25 × 5,000,000

(3) CNN

we have 

$W_1 = (900 \times 100)$

↓ Total Input

| units

So 900 is equal to 30×30 inputs as low & 100 is the number of units in first layer.

④ Lasso performs shrinkage as ridge, but additional lasso shrinks all the p predictors in the final model to zero

This is not the case in ridge where it wont set the predictors to zero especially when $(\lambda = \infty)$

Lasso has β_j coefficients as $|\beta_j|$ in terms of penalty. In all together Lasso performs variable selection. As a result, it is easier to interpret model generated in Lasso compared to ridge. Thus Lasso yields sparse model. i.e. models that involve only a subset of variables.

⑤ While performing regression, bagging we constructed B regression trees using B bootstrapped training sets and in the end we averaged resulting predictions for new sample observation.

But for considering classification for a given observation we record the closest predicted by each of B trees and the maximum, overall prediction is the most occurring

class in the B predictors

(6) Bagging does not lead to substantial reduction in variance this is because predictions we make ~~from~~ bagger trees are highly correlated.

This problem is overcome by random forests by forcing each split to consider only a subset of predictors. Thus on an average $\left(\frac{p-m}{p}\right)$ of splits will not even consider strong predictors and thus all the predictors have more chance

Thus de-correlating by making average of resulting trees less varying & more reliable.

~~number of splits~~ → $p = 225$, $m = 15$

$$\Rightarrow \frac{p-m}{p} = \frac{225-15}{225} = 0.955$$

7) given $d = 5$

k^{th} principal component PC_1 explains 35%

PC_2 explains 25%

PC_3 explains 20%

PC_4 " 15%

PC_5 " 5%

The 80% variance can be explained using first three PC's

$$PC_1 + PC_2 + PC_3 = 35 + 25 + 20$$
$$= 80$$

standard deviation of support vector of the given values

$$PC_1 = \sqrt{35} = 5.91$$

$$PC_2 = \sqrt{25} = 5$$

$$PC_3 = \sqrt{20} = 4.47$$

(8) $K = 2$, $c_1 = \{1, 2, 3, 4\}$, $c_2 = \{-9, -8, -7, -6\}$

avg $\left\{\frac{1+2+3+4}{4}\right\}$ avg $\left\{\frac{-9-8-7-6}{4}\right\}$

$(2.5, -7.5)$ are the centroids of clusters

we have to check

$$2 \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$$

reaches the optimum value

-> The results we get will rely on the initial (random) cluster assignment of each observation

Part - II

① λ plays the main role for determining the smoothness of fitting function via application of a penalty term to a loss function

The tuning parameter λ controls the flexibility of the smoothing splines and thus the effective degree of freedom.

→ As 'λ' goes ~~from~~ 0 to '∞', df_λ effective degree of freedom decreases from n to 2.

df_λ, as a measure of flexibility of smoothing splines, the higher it is, the more flexible

$$\hat{g}_\lambda = S_\lambda y$$

$\hat{g}_\lambda$ = solution to the particular choice of λ.

→ In fitting the smoothing splines, we need to choose the value of λ. In other words, we can find the value of λ that make the cross-validated RSS as small as possible.

② Slack variables that allow individual observations to be on wrong side of the margin (or) the hyperplane

we now see the three cases

i) $\varepsilon_i = 0$, the i^{th} observation is on correct side of the margin

ii) $\varepsilon_i > 0$, then i^{th} observation is on wrong side of the ~~neg~~ margin and we can say that i^{th} observation has violated the margin.

iii) $\varepsilon_i > 1$, then it is on wrong side of hyperplane

→ As the budget $\frac{C}{2}$ increases, we become tolerant of violations to the margin and so the margin will ~~be~~ be widen.

→ when $\frac{C}{2}$ is larger the margin allows more violations to it. so that we can have many support vectors.

→ 'c' also amounts to fitting the less hard data & obtaining a classifier that is potentially more biased but have lower variance.

③ A node within a decission tree contains 100 observations split evenly between two class types C_1 & C_2.

The ~~Gini~~ Gini Index

$$G = \sum_{k=1}^{K} \hat{P}_{mk}(1 - \hat{P}_{mk})$$

Alternative to this is entropy

$$D = -\sum_{k=1}^{K} \hat{P}_{mk} \log \hat{P}_{mk}$$

Since $0 \log_2 0 = 0$

→ Gini index takes on a small value if all the $\hat{P}_{mk}$ close to zero

→ Entropy will be zero, if $\hat{P}_{mk}$ are all near to 0.

Thus, entropy will take small value if m^{th} node is pure

→ When building classification tree, Both these will evaluate the quality of particular split, As these two approaches are more sensitive to node purity than the classification error rate.

④ Hierarchal clustering process

We know euclidian dist. for $[(x,y)\ (a,b)]$

$$\rightarrow \sqrt{(x-a)^2+(y-b)^2}$$

dist. $(P_1, P_2) = \sqrt{(4-6)^2 + (5-13)^2} = 8.246$

$(P_1, P_3) = \sqrt{(4-1)^2 + (5-1)^2} = 5$

$(P_2, P_3) = \sqrt{(6-1)^2 + (13-1)^2} = 13$

We perform dist. dissimilarity matrix

	P_1	P_2	P_3
P_1	0		
P_2	8.24	0	
P_3	5	13	0

so, the cluster 1 is formed between P_1 & P_3

dist matrix for computing linkage

	(1,3)	2
(1,3)	0	
2	13	0

$\max [dist(P_1 P_3), P_2]$

$dist(P_1, P_2)\ (P_3, P_2)$

$\max[13, 8.24]$

$\Rightarrow 13$

Single linkage

$\min [dist(P_1, P_3, P_2)] \Rightarrow \min[(P_1 P_2)(P_3, 2)]$

$\Rightarrow \min[13, 8.24]$

$\Rightarrow 8.24$

centroid for cluster 1 $\Rightarrow \left(\frac{1+4}{2}\right), \left(\frac{1+5}{2}\right)$

(P_1, P_3)

$(2.5, 3)$

Part - III

(1) ~~In~~ Marginal Maximum margin classifier:-

a) In this we can calculate the distance of each training observation from given hyperplane, the maximum margin hyperplane is farthest from the training observations, the smallest distance is margin

b) The maximum margin hyperplane is the ~~largest~~ minimum distance from training data, with understanding this we can classify the data according to which side it is

Support Vector Classifier :

a.) ~~In~~ support vector classifier instead of seeking the largest possible margin so that all the observations lie on the correct side of plane, we allow some observations to be on the wrong side of the hyper plane

In this case a small subset of observations called support vectors play crucial role in determining hyperplane.

This increase the accuracy of prediction that lie farther away from hyperplane.

The maximum margin classifier depends upon small set of points called support vectors present in the data. The support vectors movement or the change in support vector leads to change in movement of maximal margin hyperplane.

In svc the points that lie on the ~~wrong~~ correct side do not play any effect on the hyperplane, it is the points that lie on the margin and wrong side that effect the hyperplane. These are called support vectors.
svc is ~~not~~ robust.

→ svc has lower bias higher variance, the bias variance is controlled by tuning parameter 'c'

when 'c' is large it will have low variance and high bias

whe 'c' is small it has low bias but high variance.

Bonus Question

1) 301

2) Professor Geordie Williamson

3) Capsule Networks

4) Under 6 years old

5) IBM

6) PYTHIA

7) an IKEA chair