Data Preperation and Analysis

Final Exam

PART-II

1. λ is the tuning parameter which controls the effective degree of freedom of smoothing spline. As λ varies between 0 to infinity, the value of degree of freedom varies from n to 2.

To fit a smooth curve to a dataset, we need to find a function $g(x)$ such that, $RSS = \sum_{i=1}^{n} (y_i - g(x_i))^2$ is minimum

One way to find such a smooth function is to minimize,

$$\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt.$$

Scanned with CamScanner

where λ is a nonnegative turing parameter. The function that minimizes this is called as smoothing spline. The first part is a loss function and the second term is called penalty.

$\lambda \rightarrow$ larger the value of λ, smoother the g as well.

→ when $\lambda = 0$ the given model will be very flexible and interpretable the training data.

→ when $\lambda \rightarrow \infty$, the model corresponds to least squares linear regression.

λ controls the effective degree of freedom (df_λ)

→ As λ increases from 0 to ∞, the effective degree of freedom (df_λ) decreases from n to 2.

where, n is number of parameters in smoothing spline and hence n is normal degree of freedom.

Scanned with CamScanner

λ can be chosen by cross validation

The way RSS is calculated is

$$RSS(\lambda) = \sum_{i=1}^{n} (y_i - \hat{g}_\lambda^{(-i)}(x_i))^2 = \sum_{i=1}^{n} \left[\frac{y_i - \hat{g}_\lambda(x_i)}{1 - \{S_\lambda\}_{ii}} \right]^2$$

The matrix S_λ can be completed as

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

The effective degree of freedom for the smoothing spline is given as:

$$df_\lambda = \sum_{i=1}^{n} (S_\lambda)_{ii}$$

The maximum value that df_λ can take is n, i.e, the number of parameters in smoothing spline.

The minimum value that df_λ can take is 2 and it represents a linear model.

PART-II

2. Slack variables that allows individual observations to be on the wrong side of the margin or the hyperplane.

Scanned with CamScanner

- when $\varepsilon_i = 0$, i^{th} observation is on correct side of margin.
- when $\varepsilon_i > 0$, ith observation is on wrong side of margin and we can say it has violated the margin.
- when $\varepsilon_i > 1$, ith observation is on wrong side of hyperplane as budget C increase we become tolerant of violantion to the violation to margin and so margin will be widen
- when C is larger, margin allows more violation to it, so we can have many support vectors.
- C also amounts to fitting the less hard data and obtaining a classifier that is potentially more biased but lower variance.

Scanned with CamScanner

PART-II

3. As per the question decision tree $n=100$, split - C_1, C_2

100 observation in a node split evenly between the 2 classes C_1 and C_2

So C_1 has 50 observations and C_2 has 50 observation.

→ Gini coefficient $= 1 - \sum [P(j)]^2$ where $P(j)$ is the probability of getting class j observation thus entropy value $= 1 - [(50/100)^2 + (50/100)^2] = 1 - 1/4 - 1/4$

$= 1 - 1/4 - 1/4 = 1/2 = 0.5$

→ Entropy value $= -\sum P(j) * \log_2(P(j))$

where $P(j)$ is probability of getting class j observation thus entropy value

$= -[(50/100) * \log_2(50/100) + (50/100) * \log_2(50/100)] = 0.5 + 0.5 = 1$

Scanned with CamScanner

Now if optimal split occur into 2 leaf nodes then both the nodes will be pure i.e; both will have only observations belonging to single class only.

for leaf node with class C_1.

thus gini coefficient $= 1 - (50/50)^2 - (0/50)^2$

$= 1 - 1 - 0$

$= 0$ (Ans)

Entropy value $= -[(50/50) \log_2(50/50) + (0/50) * \log_2(0/50)]$

$= -[\log_2(1) + 0 \log_2 0] = 0.$

for leaf node with class C_2

thus gini coefficient =

$= 1 - (0/50)^2 - (50/50)^2$

$= 1 - 0 - 1 = 0$ (Ans)

Entropy value $= -[(0/50) \log_2(0/50) + (50/50) * \log_2(50/50)] = -[0 \log_2 0 + \log_2 1] = 0$

$= 0.$

Scanned with CamScanner

PART-II

4. Hierarchial clustering process with points $P1(4,5)$ and $P2(6,13)$, $P_3(1,1)$

Euclidean distance $f[(x,y),(a,b)] = \sqrt{(x-a)^2 + (x-b)^2}$

Scanned with CamScanner

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

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

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

we perform dissimilarity matrix

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

So cluster 1 is formed between P_1 & P_3. Distance Matrix for complete linkage.

$$\begin{array}{cc} & \\ 1.5 & \\ 2 & \end{array} \begin{bmatrix} 1.5 & 2 \\ 0 & \\ 13 & 0 \end{bmatrix}$$

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

$\text{dist}(P_1, P_2)(P_2, P_1)$

$\max(13, 8.24) \Rightarrow 13.$

Single linkage: $\min[\text{dis}(P_1, P_2, P_3)] = \min[(P_1, P_2), (P_3, P_2)]$

$= \min(13, 8.24)$

$= 8.24$

centroid for cluster (P_1, P_3) $= \left(\frac{1+4}{2}\right), \left(\frac{1+5}{2}\right)$

$= (2.5, 3)$

Scanned with CamScanner

PART-III

1. Maximal-Margin Classifier:

Maximum margin hyperplane is the farthest from the training observation which we compute the distance of each training observation from given hyperplane. The smallest distance is margin. Based on this we can classify the data, on which side it lies. The distance between a line and closest data points of the trained data is known as maximum-marginal classifier.

The maximal margin classifier a test observation x^* based on sign of

$$f(x^*) = \beta_0 + \beta_1 x_1^* + \ldots\ldots + \beta_p x_p^*$$

where $\beta_0, \beta_1 \ldots\ldots \beta_p$ are coefficient of maximal margin hyperplane. The maximal margin hyperplane represents mid-line of gap

Scanned with CamScanner

between the two classes.
There are relatively few training observations which are constituted of p-dimensional vectors, are those that would cause the maximal hyperplane to move if they were moved in some dimension. The maximal margin hyperplane is the solution to the optimization problem of choosing $\beta_0 \ldots \beta_p$ to maximize M such that.

$$\sum_{j=1}^{p} \beta_j^2 = 1$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}) \geq M$$

Support vector classifier

The margin in support vector machine is computed as the perpendicular distance from line to only closest points of trained data. Support vector classifier separates points that are relevant in defining the line and in construction of classifier.

Scanned with CamScanner

SVC is generalization of maximum margin classifier to non separable case is known as SVC. A hyperplane is line that splits the input variable space. The role of separating hyperplane is to accurately classify all the learning examples for splitting the line.
Optimal hyperplane is used in optimization estimation of enhance the generalized ability of support vector machine.
i) Non linear machine learning produces classifiers that have a lower bias but higher variance. Support vector machine algorithm has low bias and high variance. The bias can be increased by increased the parameters that affects the numbers of violations of margin that are allowed in data to be trained.

$$\sum_{j=1}^{p} \beta_j^2 = 1$$

$$y_i(\beta_0 + \beta_1 x_i + \ldots \beta_p x_{ip}) \geq M(1-\epsilon_i)$$

$$\epsilon_i \geq 0 \qquad \sum_{i=1}^{n} \epsilon_i \leq C$$

Scanned with CamScanner

→ The maximal marginal classifier depends on smaller set of points called support present in data.

→ In SVC, effects of point on right side of plane is nill, but effects of points on wrong side of plane is more. Hence they are called support vectors.

→ SVC is robust in nature, because SVC works with inseparable class data and hyperplane depends upon support vectors.

→ In SVC, when c is large, it will have low variance and high bias, when c is small it will have bias & high variance.

Scanned with CamScanner

PART-I

1. 3-cut points in our data, we fit piece wise polynomial for regression -

→ The first constraint is that the fitted curve must be continuous at the knots.

→ Second constraint first derivative of the peice wise polynomials has to be continuous.

→ Third constraint second derivative of the peicewise polynomials has to be continuous.

→ To get a cubic spline with K-knots uses K+4 degree of freedom

⇒ 3-knots

Degrees of freedom = 4+3 = 7.

Scanned with CamScanner

PART-I

2. d = 250 features

N = 5,000,000 observations.

we have d = 250 features

length of covariance matrix = 250×250.

number of eigen vectors = 250×250

number of eigen values = 250×1

Now, we take 10% of eigen vector and values,

∴ dimensions of projected data

sample matrix = 25×5,000,000.

Scanned with CamScanner

PART-I

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

* Unlike 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. Overall lasso performs variable selection. As a result, it is easier to interept model generated in lasso compared to ridge.

Hence, lasso yields sparse model, that is model that involve only a subset of variables.

5. when performing regression, the bagging we constructed B regression trees using B bootstrapped training sets and finally we averaged resulting

Scanned with CamScanner

predictions for new sample observations. But considering classification for a given observation we record the class predicted by each of B trees and maximum, overall prediction is most commonly occurring class among B predictors

6. p = 225 predictors

∴ $m = \sqrt{p} = 15$.

In order to find no. of split for given tree in random forest will not consider strong predictor.

∴ On average $= \frac{(p-m)}{p}$

$= \frac{225-15}{225} = 0.9333\ldots$

∴ On average, 0.93 of splits will not even consider by random forest we can think of this process as decorrelating the trees, thereby making the average of resulting trees

Scanned with CamScanner

resulting trees less variable and hence more reliable.

7. d=5 columns, variance -100

eigen values [35, 25, 20, 15, 5]

∴ here $\lambda_1 = 35$, $\lambda_2 = 25$, $\lambda_3 = 20$, $\lambda_4 = 15$
$\lambda_5 = 5$.

also $\sum_{i=1}^{5} \lambda_i = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5 = 100$

So. $\frac{\lambda_1}{\Sigma \lambda_i}$ explain the variance of 1st P.C

$\frac{\lambda_1 + \lambda_2}{\Sigma \lambda_i}$ explain variance of 1st & 2nd P.C

Since, we want to reduce dimension with 80%. So we need to find pairs where sum is 80

So, $\lambda_1 + \lambda_2 + \lambda_3 = 35 + 25 + 20 = 80$

$\lambda_1 + \lambda_2 + \lambda_4 + \lambda_5 = 35 + 25 + 15 + 5 = 80$.

So we have pairs

$[\lambda_1 + \lambda_2 + \lambda_3, \lambda_4, \lambda_5]$ and $[\lambda_1 + \lambda_2 + \lambda_4 + \lambda_5, \lambda_3]$

Scanned with CamScanner

$[80, 15, 5]_{1\times3}$ and $[20, 20]_{1\times2}$

Here 2 dimension are reduced   here 3 dimensions and reduced

Now,

$$\sigma_1 = \left\{\frac{1}{3}\left(80 - \frac{100}{3}\right)^2 + \left(15 - \frac{100}{3}\right)^2 + \left(5 - \frac{100}{3}\right)^2\right\}^{1/2}$$

$$\sigma_1 = \left[\frac{1}{3}\left\{\left(\frac{140}{3}\right)^2 + \left(\frac{55}{3}\right)^2 + \left(\frac{85}{3}\right)^2\right\}\right]^{1/2}$$

$$\sigma_1 = \frac{1}{3\sqrt{3}}\left[(140)^2 + (55)^2 + (85)^2\right]^{1/2}$$

$$\sigma_1 = \frac{1}{3\sqrt{3}}\left[19600 + 3025 + 7225\right]^{1/2}$$

$$\sigma_1 = \frac{1}{3\sqrt{3}}\left[29850\right]^{1/2} = \frac{172.771}{3\sqrt{3}}$$

$$\sigma_1 = \frac{57.520}{\sqrt{3}} = 33.24$$

Similarly

$$\sigma_2 = \left[\frac{1}{3}\left\{\left(80 - \frac{100}{2}\right)^2 + \left(20 - \frac{100}{2}\right)^2\right\}\right]^{1/2}$$

$$= \left[\frac{1}{3}\left\{\left(\frac{60}{2}\right)^2 + \left(\frac{60}{2}\right)^2\right\}\right]^{1/2}$$

Scanned with CamScanner

$$\sigma_2 = \frac{1}{2\sqrt{3}} \left[(60)^2+(60)^2\right]^{1/2}$$

$$\sigma_2 = \frac{1}{2\sqrt{3}} \left[2(60)^2\right]^{1/2}$$

$$\sigma_2 = \frac{\sqrt{2}\cdot 60}{2\cdot\sqrt{3}}$$

$$\sigma_2 = \frac{30\sqrt{2}}{\sqrt{3}} = 24.434$$

So Analysis able to reduce 2 dim in $[80, 15, 5]$

reduce 3 dim in $[80, 20]$

Standard deviation of $[80, 15, 5]$ $\sigma_1 = 33.24$

Standard deviation of $[80, 20]$ $\sigma_2 = 24.494$

8. $K=2$,

$C_1 = \{1, 2, 3, 4\}$

$C_2 = \{-9, -8, -7, -6\}$

centroid of clusters are

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

$= (2.5, -7.5)$

Scanned with CamScanner

we have to check

$$2 \sum_{i \in C_k} \sum_{j=k}^{P} (x_{ij} - \bar{x}_{kj})^2 \text{ reaches}$$

optimum values.

- Result obtained will depend on the initial (random) cluster assignment of each observation.

Scanned with CamScanner

Lucky 7

1. 2662
2. Kazhdan – Lusztig polynomial. Knot Theory and Representation theory
3. Capsule neural network
4. Under 18 years of age
5. IBM
6. PYTHIA
7. IKEA Chair

Scanned with CamScanner

Part - I

3. First layer of CNN – 100 units

gray scale images – 30×30 pixel

size of weight matrix w_1,

(Image size)
$N \times N = 30\times30$.

F filter size

$F \times F = 10\times10$ (100 units)

$i/p = 30\times30 = 900$.

weight $= (N-F+1)\times(N-F+1)$

$= (30-10+1)\times(30-10+1)$

$= 21\times21$
Rows columns.

Scanned with CamScanner