Midterm II Review

Fall 2021

Lecture Overview

1. **Midterm Structure**
2. Naïve Bayesian Classifier
3. Fisher LDA
4. Perceptron
5. Trees with MDL
6. kNN classification

Midterm Description

- Total comes to 90 pts. However, max you can earn is capped at 70.
- On one sheet exam for graduate and for undergraduate – however questions are marked as follows:
  - UG (e.g., 'problem 1UG') – for both graduate and undergraduate
  - G – (e.g. 'problem 4G') for graduate

  Please pay attention and avoid confusion.
- 3 questions UG + 1 question G
- Computationally intensive - you should use Wolfram alpha or calculator.
- You MUST give ALL intermediate results whenever asked for. No intermediate results no credit

Midterm Description - continued

- Credit for each question clearly marked. Partial credit possible, also marked.
  - Credit is based on undergraduate credit
  - Undergrad credit for a UG question is given in brackets e.g. [30] means that a question gives undergraduate student 30 pts.
  - Graduate credit is given in curly brackets. For UG questions it is given as a multiplier to apply to undergrad credit, e.g. [30]{2/3} means that grad students get for this question 30*2/3=20 points. Same multiplier applies to all partial points
  - Grad credit for grad only questions (G) credit is given in curly brackets {25} means grad students get 25 pts for this question

Midterm

1. [35pts]{5/7} – Fisher LDA/Perceptron. Similar to Fisher LDA/perceptron problems in HW 5A. Given a data table and a new record (not in the table). Compute Fisher LDA classifier/Perceptron classifier from the table and the classify the new record using this learned classifier
2. [30pts]{2/3} – Naïve Bayesian Classifier. Similar to 5-7 in HW 5UG . Given a data table and a new record (not in the table). Compute Naïve Bayesian Classifier from the table a classify the new record using this learned classifier.
3. [25pts]{1.0} – MDL. Similar to MDL problem in HW 4. Given two decision trees (i.e. trees itself, the set of attributes, description of splits and classes of leafs) with specified training error. Need to find which tree is better w.r.t MDL

Midterm - continued

4. [20 pts] – kNN classifier. Similar to HW5G. Nearest neighbor classifier. Given a training set and a new datapoint. Classify the new datapoint using kNN classifier.

Lecture Overview

1. Midterm Structure
2. **Naïve Bayesian Classifier**
3. Fisher LDA
4. Perceptron
5. Trees with MDL
6. kNN classification

Parameters of Naïve Bayesian Classifier

- Consider the data set shown in the table. Estimate the conditional probabilities (using pmf/Bernoulli) for $Pr(A|+)$; $Pr(B|+)$; $Pr(C|+)$; $Pr(A|-)$; $Pr(B|-)$ and $Pr(C|-)$.
- Answer

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	−
3	0	1	1	−
4	0	1	1	−
5	0	0	1	+
6	1	0	1	+
7	1	0	1	−
8	1	0	1	−
9	1	1	1	+
10	1	0	1	+

$$Pr(A = 1|-) = 2/5 = 0.4;$$
$$Pr(B = 1|-) = 2/5 = 0.4;$$
$$Pr(C = 1|-) = 1;$$
$$Pr(A = 0|-) = 3/5 = 0.6;$$
$$Pr(B = 0|-) = 3/5 = 0.6;$$
$$Pr(C = 0|-) = 0;$$
$$Pr(A = 1|+) = 3/5 = 0.6;$$
$$Pr(B = 1|+) = 1/5 = 0.2;$$
$$Pr(C = 1|+) = 4/5 = 0.8;$$
$$Pr(A = 0|+) = 2/5 = 0.4;$$
$$Pr(B = 0|+) = 4/5 = 0.8;$$
$$Pr(C = 0|+) = 1/5 = 0.2.$$

Naïve Bayesian Classifier – Record Classification

- Use the estimate of conditional probabilities given in the previous question to predict the class label for a test sample (A = 0;B = 1;C = 0) using the naive Bayes approach.

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	−
3	0	1	1	−
4	0	1	1	−
5	0	0	1	+
6	1	0	1	+
7	1	0	1	−
8	1	0	1	−
9	1	1	1	+
10	1	0	1	+

- Answer

Let $\Pr(A = 0; B = 1; C = 0) = K$. Then

$$\Pr(+|A = 0; B = 1; C = 0) = \frac{\Pr(A = 0; B = 1; C = 0|+)\Pr(+)}{P(A = 0; B = 1; C = 0)}$$

$$= \frac{P(A = 0|+)P(B = 1|+)P(C = 0|+)\Pr(+)}{K}$$

$$= \frac{0.4 \times 0.2 \times 0.2 \times 0.5}{K} = \frac{0.008}{K}$$

$$\Pr(-|A = 0; B = 1; C = 0) = \frac{\Pr(A = 0; B = 1; C = 0|-)\Pr(-)}{P(A = 0; B = 1; C = 0)}$$

$$= \frac{P(A = 0|-)P(B = 1|-)P(C = 0|-)\Pr(-)}{K}$$

$$= \frac{0.6 \times 0.4 \times 0 \times 0.5}{K} = \frac{0}{K}$$

The class label should be + .

Parameters of Naïve Bayes: m-estimate

- Estimate the conditional probabilities using the m −estimate approach, with $p = \frac{1}{2}$ and $m = 4$.
- Answer

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	−
3	0	1	1	−
4	0	1	1	−
5	0	0	1	+
6	1	0	1	+
7	1	0	1	−
8	1	0	1	−
9	1	1	1	+
10	1	0	1	+

$$P(A = 0|+) = \frac{2+2}{5+4} = 4/9;$$

$$P(A = 0|-) = \frac{3+2}{5+4} = 5/9;$$

$$P(B = 1|+) = \frac{1+2}{5+4} = 3/9;$$

$$P(B = 1|-) = \frac{2+2}{5+4} = 4/9;$$

$$P(C = 0|+) = \frac{1+2}{5+4} = 3/9;$$

$$P(C = 0|-) = \frac{0+2}{5+4} = 2/9.$$

Naïve Bayes w/m-estimate – Record Classification

- Use the estimate of conditional probabilities given in the previous question to predict the class label for a test sample (A = 0;B = 1;C = 0) using the naive Bayes approach.

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	−
3	0	1	1	−
4	0	1	1	−
5	0	0	1	+
6	1	0	1	+
7	1	0	1	−
8	1	0	1	−
9	1	1	1	+
10	1	0	1	+

- Answer

Let $\Pr(A = 0; B = 1; C = 0) = K$. Then

$$\Pr(+|A = 0; B = 1; C = 0) = \frac{\Pr(A = 0; B = 1; C = 0|+)\Pr(+)}{P(A = 0; B = 1; C = 0)}$$

$$= \frac{P(A = 0|+)P(B = 1|+)P(C = 0|+)\Pr(+)}{K}$$

$$= \frac{(4/9)(3/9)(3/9)0.5}{K} = \frac{0.0247}{K}$$

$$\Pr(-|A = 0; B = 1; C = 0) = \frac{\Pr(A = 0; B = 1; C = 0|-)\Pr(-)}{P(A = 0; B = 1; C = 0)}$$

$$= \frac{P(A = 0|-)P(B = 1|-)P(C = 0|-)\Pr(-)}{K}$$

$$= \frac{(5/9)(4/9)(2/9)0.5}{K} = \frac{0.0274}{K}$$

The class label should be - .

Lecture Overview

1. Midterm Structure
2. Naïve Bayesian Classifier
3. **Fisher LDA**
4. Perceptron
5. Trees with MDL
6. kNN classification

Classify with Fisher LDA

- Given the following training data set (table on the left)

  Find Fisher's normal vector $\overline{w}$ and intercept sep (separation point) that define the classification plane $[\overline{w}: \text{sep}]$.

#	Data Points (A_1, A_2)	class
1	(2.5, 1.0)	0
2	(2.0, 2.15)	0
3	(4.0, 2.9)	1
4	(3.6, 4.0)	1

1. Compute means μ_1 and μ_2, and the between-class scatter matrix S_B
2. Compute SC_1 and SC_2, the within-class scatter matrices and their sum S_W
3. Find the optimal vector $\overline{w}$ that discriminates between the classes
4. Having found discriminant vector $\overline{w}$, find the point sep on $\overline{w}$ that best separates the two classes.
5. Classify the point $\bar{x} = \begin{pmatrix} 3.8 \\ 5 \end{pmatrix}$

Transposing data:

#	1	2	3	4
A_1	2.5	2.0	4.0	3.6
A_2	1.0	2.15	2.9	4.0
class	0	0	1	1

$class\ data\ X_1$ $class\ data\ X_2$

Classify with Fisher LDA - continued

- $X_1 = \{\{2.5,2\}, \{1,2.15\}\}$
- $X_2 = \{\{4,3.6\}, \{2.9,4\}\}$
- $\mu_1 = (2.25,1.575)$ and $\mu_2 = (3.8,3.45)$
- $X_1^C = \{\{0.25,-0.25\}, \{-0.575,0.575\}\}$
- $X_2^C = \{\{0.2,-0.2\}, \{-0.55,0.55\}\}$
- $\mu_2 - \mu_1 = (1.55, 1.875)$
- $S_B = (\{\{1.55, 1.875\}\}^T).\{\{1.55, 1.875\}\} = \{\{2.40,2.91\},\{2.91,3.52\}\}$
- $SC_1 = \{\{0.25,-0.25\}, \{-0.575,0.575\}\}.(\{\{0.25,-0.25\},\{-0.575,0.575\}\}^T)$
  $= \{\{0.125,-0.288\},\{\{-0.288,0.661\}\}$
- $SC_2 = \{\{0.2,-0.2\}, \{-0.55,0.55\}\}.\left(\{\{0.2,-0.2\}, \{-0.55,0.55\}\}^T\right)$
  $= \{\{0.08,-0.22\},\{-0.22,0.605\}\}$
- $S_W = SC_1 + SC_2 = \{\{0.205,-0.508\},\{-0.508,1.266\}\}$
- $S_W^{-1} = \{\{863:574,346:521\},\{346.521,139.836\}\}$
- $S_W^{-1} S_B = \{\{3080.955,3732.756\},\{1238.574,1500.600\}\}$

Classify with Fisher LDA - continued

- Largest eigenvalue of $S_W^{-1}S_B$ is $\lambda_{max} = 4581.56$
- Its normalized eigenvector is $\overline{w} = \begin{pmatrix} 0.927832 \\ 0.372997 \end{pmatrix}$
- Class 1: $\mu_{\overline{w}}^1 = 2.675\overline{w}$
- Class 2: $\mu_{\overline{w}}^2 = 4.812\overline{w}$
- % of variance of classes is approximately equal so classifying point is:
  $sep = \frac{2.675+4.812}{2}\overline{w} = 3.743\overline{w}$
- $proj_{\overline{w}}\bar{x} = \begin{pmatrix} 0.927832 \\ 0.372997 \end{pmatrix}^T \begin{pmatrix} 3.8 \\ 5 \end{pmatrix} \overline{w} = 5.39\overline{w}$

  so the new point is classified as class 2

Lecture Overview

1. Midterm Structure
2. Naïve Bayesian Classifier
3. Fisher LDA
4. **Perceptron**
5. Trees with MDL
6. kNN classification

Perceptron Problem

Apply the perceptron learning algorithm for the following pattern set until convergence. Start with 0-weight vector

$$\overline{w} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

For simplicity of computation use $\eta = 1$. Apply the learning algorithm to data points in the given order cyclically. For each step of perceptron learning write down the classification result of a datapoint with the current weight vector, indicator if update is needed, and computation of vector update if necessary. The dataset consist of datapoints a, b, c, d:

	X_1	X_2	X_3	Y
$a.$	4	3	6	-1
$b.$	2	-2	3	1
$c.$	1	0	-3	1
$d.$	4	2	3	-1

Perceptron - Solution

- Solution: $w_0 = \overline{w} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$

step	data point	$\text{sgn}(\overline{w}_i \bullet x_i)$	true class	adjustment	$\overline{w}_{i+1}$
1	$a = (4,3,6,1)^T$	$+$	$-$	$-(4,3,6,1)^T$	$(-4,-3,-6,0)^T$
2	$b = (2,-2,3,1)^T$	$-$	$+$	$(2,-2,3,1)^T$	$(-2,-5,-3,1)^T$
3	$c = (1,0,-3,1)^T$	$+$	$+$	no	$(-2,-5,-3,1)^T$
4	$d = (4,2,3,1)^T$	$-$	$-$	no	$(-2,-5,-3,1)^T$
5	$a = (4,3,6,1)^T$	$-$	$-$	no	$(-2,-5,-3,1)^T$
6	$b = (2,-2,3,1)^T$	$-$	$+$	$(2,-2,3,1)^T$	$(0,-7,0,2)^T$
7	$c = (1,0,-3,1)^T$	$+$	$+$	no	$(0,-7,0,2)^T$
8	$d = (4,2,3,1)^T$	$-$	$-$	no	$(0,-7,0,2)^T$
9	$a = (4,3,6,1)^T$	$-$	$-$	no	$(0,-7,0,2)^T$
10	$b = (2,-2,3,1)^T$	$+$	$+$	no	$(0,-7,0,2)^T$

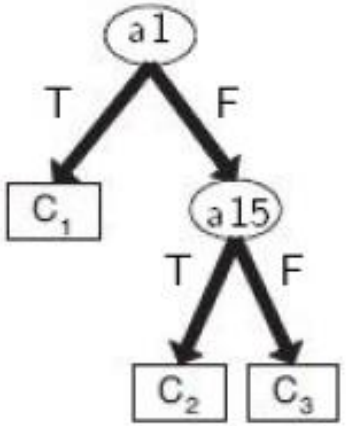
- So vector $\vec{w}_7 = \begin{pmatrix} 0 \\ -7 \\ 0 \\ 2 \end{pmatrix}$ classifies points correctly

Lecture Overview

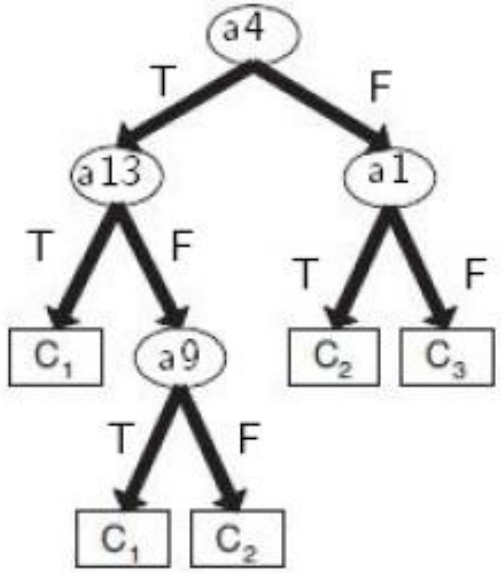
1. Midterm Structure
2. Naïve Bayesian Classifier
3. Fisher LDA
4. **Trees with MDL**
5. kNN classification

Compare 2 DT Classifiers w.r.t MDL

- Consider the decision trees shown in Figure 4.3. Assume they are generated from a data set that contains 16 binary attributes and 3 classes, C1, C2, and C3. Let the data contain total of n data points classifed by the trees.
  a. Compute the MDL bound for each decision tree as a function of $n =$ sample size and $\delta =$ confidence level.
  b. Which tree is better for $n = 200$ and confidence 0.99 ($\delta = 0.01$)



(a) Decision tree with 7 errors



(b) Decision tree with 4 errors

Compare 2 DT Classifiers w.r.t MDL

a. Solution:

- Fix size-first description of DT for binary classification using - encoding that is known to be prefix free: all features are numbered from 2 to *k* + 1 and 'leaf' designation is treated as a feature #1 that has class as 'domain values' (i.e. {1, 2,…,k}).
- number of features in the tree (+1 for class)
- size of the sample
- maximum branching degree
- number of classes
- the following sequence of nodes is given in BFS order of walking the decision tree:
  - number of children of the node (since there are at least 2 children 1 stands for no children),
  - the number of a feature used in the split,
  - domain value used in a split as the number of example in use.

Compare 2 DT Classifiers w.r.t MDL – cont.

- Gammais used to encode every number in the tree description, i.e. k bits in binary representation of a number is transformed into 2k+1 bits of gamma encoding.

1. Left tree: It has 5 nodes so it needs $2\lceil \log_2 5 \rceil + 1 = 7$ bits (nodes encoding=ne)

- Number of features +1 (for class): $2\lceil \log(15+1) \rceil + 1 = 9$ (feature enc.=fe)
- Sample size requires $2\log\lceil 200 + 3 \rceil + 1 = 17$ bits (3 for classes, domain size =ds)
- Branching degree is 2 so it requires $2 \times 1 + 1 = 3$ bits (branching degree =bd)
- Number of classes is 3 so $2\lceil \log 3 \rceil + 1 = 5$ bits (number of classes = nc)
- Each node requires 9 bits (feature code)+3 bits (number of children)+17bits(domain value)=29bits (node size =ns)
- So total length for the tree is

  $7(ne) + 9(fe) + 17(ds) + 3(bd) + 5(nc) + 5 \times 29(nds \times ns) = 186$ bits

2. Right tree

- It has 8 nodes so it needs $2\lceil \log_2 8 \rceil + 1 = 9$ bits
- Number of features +1 (for class) required $2\lceil \log(13+1) \rceil + 1 = 9$
- Sample size, branching degree and number of classes is the same as in tree 1
- Hence, each node requires same number of bits as left tree -35
- So total length for the tree is $9 + 9 + 17 + 3 + 5 + 9 \times 29 = 299$

Compare 2 DT Classifiers w.r.t MDL – cont.

- Using MDL bound $L_S(h) + \sqrt{\frac{\log_2\left(\frac{2}{\delta}\right)+|h|}{2n}}$ on generalization error, we get for left tree

$\frac{7}{n} + \sqrt{\frac{\log_2\left(\frac{2}{\delta}\right)+186}{2n}}$ and $\frac{4}{n} + \sqrt{\frac{\log_2\left(\frac{2}{\delta}\right)+299}{2n}}$ for right tree.

b. Solution:

- When is left tree better that right? when

$$\frac{7}{n} + \sqrt{\frac{\log_2\left(\frac{2}{\delta}\right)+186}{2n}} < \frac{4}{n} + \sqrt{\frac{\log_2\left(\frac{2}{\delta}\right)+299}{2n}}$$

that is

$$\frac{3}{n} < \frac{\sqrt{\log_2\left(\frac{2}{\delta}\right)+186} - \sqrt{\log_2\left(\frac{2}{\delta}\right)+299}}{\sqrt{2n}}$$

- At this point we need to plug in the value of $\delta = .01$ and $n = 200$ and see what is the sign of $\frac{\sqrt{\log_2\left(\frac{2}{0.01}\right)+186} - \sqrt{\log_2\left(\frac{2}{0.01}\right)+299}}{\sqrt{2\times 200}} - \frac{3}{200}$. If it is positive then left tree is better, otherwise right tree is better. The value of the expression is -0.22, so right tree is better under these conditions

Lecture Overview

1. Midterm Structure
2. Naïve Bayesian Classifier
3. Fisher LDA
4. Trees with MDL
5. **kNN classification**

Nearest Neighbor Classifier

- Consider the one-dimensional data set

x	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
y	-	-	+	+	+	-	-	+	-	-

Classify the data point $x = 5.0$ according to its 1-, 3-, 5-, and 9-nearest neighbors (using majority vote).

- Answer:
  - 1-nearest neighbor: +,
  - 3-nearest neighbor: -
  - 5-nearest neighbor: +,
  - 9-nearest neighbor: -.

Nearest Neighbor Classifier

- Repeat the previous analysis using the distance-weighted voting approach described in Section 5:2:1.
- Answer:

Weights are $\frac{1}{d^2(x,y)}$ where x is new point and y is old point. So the nearest point at a distance 0.1 out-weigh all other points in any kNN procedure, i.e.

- 1-nearest neighbor: +,
- 3-nearest neighbor: +,
- 5-nearest neighbor: +,
- 9-nearest neighbor: +.