Data Mining & Knowledge Discovery Exam 04.03.22

SAMPLE SOLUTION

points	grade
40-	1.0
37-39.5	1.3
34-36.5	1.7
31-33.5	2.0
28-30.5	2.3
25-27.5	2.7
22-24.5	3.0
19-21.5	3.3
16-18.5	3.7
13-15.5	4.0

Problem 1: General Understanding (12 of 58 points)

How are these methods called?

Example: Minimize the sum of the edge weights of a Hamiltonian cycle of a graph: traveling salesman algorithm.

a) Maximize the variance of the data along one dimension.

One dimensional principal component analysis. (2)

b) Minimize the sum of small quadratic and large scaled absolute regression errors.

Huber's robust regression. (2)

c) Minimize the weighted sum of the squared norm of the normal vector of the discriminance hyperplane in the kernelized high–dimensional space plus the sum of the penalty terms for all data points that are not beyond the classification margin.

Support vector machine. (2)

d) Minimize the sum of distances between each data point and the average of the corresponding set of points.

c–means clustering. (2)

e) Minimize the quadratic reconstruction error after linear projection to a plane.

Two dimensional principal component analysis. (2)

f) Minimize the sum of the m smallest squared differences between the regression values and the output data.

Least trimmed squares. (2)

Problem 2: Principal Component Analysis (16 of 58 points)

Consider a data set X for which one–dimensional PCA will yield

$$X' = \{(3,4),(3,4),(-3,-4),(-3,-4)\}$$

with mean square error $e = 1$.

a) Compute the eigenvectors of the covariance matrix of X (not X').

The first eigenvector is along the data in X': The difference vector is $(3,4) - (-3,-4) = (6,8)$, with length $\sqrt{6^2+8^2} = \sqrt{36+64} = \sqrt{100} = 10$, so the eigenvector is $v_1 = (6,8)/10 = (0.6, 0.8)$. X is two–dimensional, so there is a second eigenvector, orthogonal to v_1: $v_2 = (0.8, -0.6)$. (4)

b) Compute the eigenvalues of the covariance matrix of X (not X').

The distance of each point in X' from the origin is $\sqrt{3^2+4^2} = \sqrt{9+16} = \sqrt{25} = 5$. So, the variance along the first eigenvector is $\lambda_1 = 1/(4-1) \cdot 4 \cdot 5^2 = 100/3$. The mean square error of the PCA projection is $e = 1 = (4-1)/4 \cdot \lambda_2$, so $\lambda_2 = 4/(4-1) = 4/3$. (4)

c) Find all data sets X that will lead to these results.

X' is the projection of X to the first eigenvector. So we can construct X using the points in X' plus/minus the scaled second eigenvector. To keep the main axis unchanged, the first two points have to be moved by the same distance, in opposite directions, and the same for the other two points. We can write this as $X = \{(3,4)+\alpha\cdot(0.8,-0.6),(3,4)-\alpha\cdot(0.8,-0.6),(-3,-4)+\beta\cdot(0.8,-0.6),(-3,-4)-\beta\cdot(0.8,-0.6)\}$ with $\alpha,\beta \in \mathbb{R}$. The mean square error is $e = 1/4\cdot(\alpha^2+\alpha^2+\beta^2+\beta^2) = \alpha^2/2+\beta^2/2 = 1$, so $\beta = \pm\sqrt{2-\alpha^2}$ with $\alpha \in [-\sqrt{2}, \sqrt{2}]$. Further, we can arbitrarily choose $\bar{x} = (\bar{x}_1, \bar{x}_2) \in \mathbb{R}^2$. So, the answer is all sets $X = \{(3,4)+(\bar{x}_1,\bar{x}_2)+\alpha\cdot(0.8,-0.6),(3,4)+(\bar{x}_1,\bar{x}_2)-\alpha\cdot(0.8,-0.6),(-3,-4)+(\bar{x}_1,\bar{x}_2)\pm\sqrt{2-\alpha^2}\cdot(0.8,-0.6),(-3,-4)+(\bar{x}_1,\bar{x}_2)\mp\sqrt{2-\alpha}\cdot(0.8,-0.6)\}$ with $\alpha \in [-\sqrt{2}, \sqrt{2}]$, $\bar{x}_1, \bar{x}_2 \in \mathbb{R}$. (8)

Problem 3: Regression (16 of 58 points)

Consider an n–layer MLP with one neuron per layer, no offset, transfer function

$$y = \frac{x}{\sqrt{1+x^2}}$$

All weights are initialized as 1. Training is done by minimizing the quadratic output error using gradient descent backpropagation. We use a simple training data set with only one sample: input 1 and output 0.

a) For the untrained network and the given training input, compute the output of layer k, $k = 1, \ldots, n$.

$$y_1 = \frac{1}{\sqrt{1+1^2}} = \frac{1}{\sqrt{2}}$$

$$y_2 = \frac{\frac{1}{\sqrt{2}}}{\sqrt{1+\frac{1}{2}}} = \frac{1}{\sqrt{3}}$$

$$y_3 = \frac{\frac{1}{\sqrt{3}}}{\sqrt{1+\frac{1}{3}}} = \frac{1}{\sqrt{4}}$$

$$y_k = \frac{1}{\sqrt{k+1}}$$

(4)

b) For the first training step with the given training data, compute the gradient to update the weight of the edge between layer $k-1$ and layer k, $k=1,\ldots,n$.

$$E = (y_n - 0)^2$$

$$\frac{\partial y}{\partial x} = \frac{\sqrt{1+x^2} - \frac{x^2}{\sqrt{1+x^2}}}{1+x^2} = \frac{1+x^2-x^2}{\sqrt{1+x^2}^3} = \frac{1}{\sqrt{1+x^2}^3}$$

$$\frac{\partial E}{\partial w_n} = \frac{\partial E}{\partial y_n}\cdot\frac{\partial y_n}{\partial x_n}\cdot\frac{\partial x_n}{\partial w_n} = 2y_n\cdot\frac{1}{\sqrt{1+x_n^2}^3}\cdot y_{n-1} = \frac{2y_n y_{n-1}}{\sqrt{1+y_{n-1}^2}^3} = \frac{2}{\sqrt{(n+1)\cdot n\cdot\left(\frac{n+1}{n}\right)^3}} = \frac{2n}{(n+1)^2}$$

$$\frac{\partial E}{\partial w_{n-1}} = \frac{\partial E}{\partial y_n}\cdot\frac{\partial y_n}{\partial x_n}\cdot\frac{\partial x_n}{\partial y_{n-1}}\cdot\frac{\partial y_{n-1}}{\partial x_{n-1}}\cdot\frac{\partial x_{n-1}}{\partial w_{n-1}} = \frac{\partial E}{\partial w_n}\cdot\frac{y_{n-2}}{y_{n-1}}\cdot\frac{1}{\sqrt{1+x_{n-1}^2}^3}$$

$$= \frac{2n}{(n+1)^2}\sqrt{\frac{n}{n-1}}\cdot\sqrt{\frac{n-1}{n}}^3 = \frac{2(n-1)}{(n+1)^2}$$

$$\frac{\partial E}{\partial w_{n-2}} = \frac{\partial E}{\partial y_n}\cdot\frac{\partial y_n}{\partial x_n}\cdot\frac{\partial x_n}{\partial y_{n-1}}\cdot\frac{\partial y_{n-1}}{\partial x_{n-1}}\cdot\frac{\partial x_{n-1}}{\partial y_{n-2}}\cdot\frac{\partial y_{n-2}}{\partial x_{n-2}}\cdot\frac{\partial x_{n-2}}{\partial w_{n-2}} = \frac{\partial E}{\partial w_{n-1}}\cdot\frac{y_{n-3}}{y_{n-2}}\cdot\frac{1}{\sqrt{1+x_{n-2}^2}^3}$$

$$= \frac{2(n-1)}{(n+1)^2}\sqrt{\frac{n-1}{n-2}}\cdot\sqrt{\frac{n-2}{n-1}}^3 = \frac{2(n-2)}{(n+1)^2}$$

$$\frac{\partial E}{\partial w_k} = \frac{2k}{(n+1)^2}$$

(8)

c) Which problem will occur when the number of layers becomes very large?

n large $\Rightarrow \frac{\partial E}{\partial w_k}$ very small (vanishing gradient problem) (2)

d) How can this problem be avoided?

Training layers separately, or other transfer function. (2)

Problem 4: Classification (14 of 58 points)

Consider the following data from the UK Office for National Statistics about the mortality reates by vaccination status for deaths involving COVID–19 in England, from September 18–24, 2021.

Vaccination status	Age group	Number of deaths	Population
unvaccinated	10–59	28	7167322
unvaccinated	60+	67	470189
vaccinated	10–59	34	19811645
vaccinated	60+	436	11700409

a) What is the death probability that a Naive Bayes classifier yields for a vaccinated person aged 10–59?

Vaccination status	Age group	positive	negative	total
unvaccinated	10–59	28	7167294	7167322
unvaccinated	60+	67	470122	470189
vaccinated	10–59	34	19811611	19811645
vaccinated	60+	436	11699973	11700409
sum		565	39149000	39149565

$$\frac{\frac{565}{39149565}\cdot\frac{28+34}{565}\cdot\frac{34+436}{565}}{\frac{565}{39149565}\cdot\frac{28+34}{565}\cdot\frac{34+436}{565}+\frac{39149000}{39149565}\cdot\frac{7167294+19811611}{39149000}\cdot\frac{19811611+11699973}{39149000}}\approx 2.37501\cdot 10^{-6}$$

(3)

b) What is the death probability that a Naive Bayes classifier yields for an unvaccinated person aged 10–59?

$$\frac{\frac{565}{39149565}\cdot\frac{28+34}{565}\cdot\frac{28+67}{565}}{\frac{565}{39149565}\cdot\frac{28+34}{565}\cdot\frac{28+67}{565}+\frac{39149000}{39149565}\cdot\frac{7167294+19811611}{39149000}\cdot\frac{7167294+470122}{39149000}}\approx 1.98069\cdot 10^{-6}$$

(3)

c) Compare and explain the results from (a) and (b)!

The Naive Bayes classifier indicates a death probability for a vaccinated person which is higher than for an unvaccinated person. The reason is that the features age group and vaccination status are correlated. (2)

d) From the given data (ignoring the Naive Bayes classifier), compute the conditional probability of death for a vaccinated person.

$$\frac{34 + 436}{19811645 + 11700409} \approx 14.9149 \cdot 10^{-6}$$

(2)

e) From the given data (ignoring the Naive Bayes classifier), compute the conditional probability of death for an unvaccinated person.

$$\frac{28 + 67}{7167322 + 470189} \approx 12.4386 \cdot 10^{-6}$$

(2)

f) Compare and explain the results from (d) and (e)!

The conditional death probability for a vaccinated person is higher than for an unvaccinated person. The reason is that the deaths are correlated with the vaccination status and with the age group. (2)