

패턴인식 2021 년 중간고사

학번:

성명:

학부:

[1] (5점) Set $\{(1, 1, 0), (1, 0, 1), (2, 1, 1)\}$ 가 span하는 vector space는 몇차원 공간인가?

[2] (5점) $X := [x_1 \ x_2 \ x_3]$, x_i 는 2차원 column 벡터이다. $X^T X$ 과 XX^T 를 계산하고 각각의 차원($m \times n$ 형태)을 구하시오.

[3] (5점) $A = \begin{bmatrix} 1 & 2 & -1 \\ -2 & -4 & 2 \end{bmatrix}$ 의 $C(A)$, $N(A)$ 를 구하고 각 벡터공간의 차원과 basis를 제시하시오.

[4] (5 점) A 가 $m \times n$ ($m \neq n$) matrix 일 때, Singular value decomposition 을 하면 $A = U\Sigma V^*$ 로 나타낼 수 있다. 여기서 Σ 는 일부의 대각 요소에 singular value 가 위치하고 나머지 요소는 모두 0 인 $m \times n$ matrix 이다. $Ax=b$ 이 해가 무수히 많을 때 $\|x\|$ 이 가장 적은 해를 구하시오.

[5] (5 점) Kullback – Leibler divergence $D_{p(X,Y)||p(X)p(Y)} = 0$ 이면 Mutual information $I(X,Y)$ 의 값은 얼마인가? 왜 그렇게 되는지 이유를 쓰시오.

[6] (5점) 다음의 확률 값이 주어져 있다.

$$P(X = 0|Y = 0) = \frac{2}{3}, P(X = 1|Y = 0) = \frac{1}{3},$$

$$P(X = 0|Y = 1) = \frac{3}{8}, P(X = 1|Y = 1) = \frac{5}{8},$$

$$P(Y = 0) = 1/5, P(Y = 1) = 4/5$$

$H(X,Y) = - \sum_x \sum_y p(x,y) \log p(x,y)$ 를 구하시오. 계산은 하지 마세요..

[7] (5점) 동전을 던졌을 때, 앞이 나오는 경우 확률변수 X 는 0이되고, 뒤가 나오는 경우는 X 가 1이된다. 앞이 나올 확률은 $1/3$, 뒤가 나올 확률은 $2/3$ 이다. 이 확률 변수 X 가 가지는 entropy는 얼마인가?

[8] (5점) $C_{p||q}(x; W) = -\sum_x \sum_k p_k(x) \log q_k(x; W)$ 에서 신경망의 입력 x 에 대해 신경망의 k 번째 노드의 출력이 $q_k(x; W)$ 이고 desired 출력이 $p_k(x)$ 이다. $C_{p||q}(x; W)$ 가 0이 되도록 W 를 학습하게 되면 어떤 결과를 가져오는가?

[9] (각 5 점) 두 class 를 분류하는 classifier 는 $y = w^T x + b$ 로 주어진다. 즉, 입력 패턴 x 에 대해 y 가 양수이면 label 을 +1 로 부여하고, y 가 음수이면 label 을 -1 로 부여한다. 정확한 label 을 부여하도록 하기 위한 w 값은 n 개의 훈련 데이터 $\{(x_i, y_i) | i = 1, \dots, n\}$ 를 이용하여 학습에 의해 구해진다. 이때, class 간의 margin(비무장지대)을 최대로 하는 Support Vector Machine(SVM)은 다음의 최적화문제를 풀어서 w 를 구한다.

$$\begin{aligned} & \underset{w}{\text{minimize}} && \frac{1}{2} w^T w \\ & \text{subject to} && y_i(w^T x_i + b) \geq 1, \quad i = 1, \dots, n \end{aligned}$$

A. Lagrangian $L(w, \alpha)$ 는 아래와 같이 주어지는데 괄호 안에 들어갈 수식은?

$$L(w, \alpha) = \frac{1}{2} w^T w - \sum_{i=1}^n \alpha_i (\quad)$$

B. KKT 4 번째 조건인 $\nabla_w L(w, \alpha) = 0$ 을 만족하는 w 를 구하시오.

C. KKT 3 번째 조건은 무엇인가? 이 조건은 support vector 가 아닌 패턴에 해당하는 α_i 가 0 이 된다는 것을 의미한다. 왜 그런가?

D. B 에서 구한 w 를 $L(w, \alpha)$ 에 대입하며 dual 함수 $L(\alpha) = -\frac{1}{2} \alpha^T Q \alpha + \alpha^T \mathbf{1}$ 를 구할 수 있고 이를 최대로 하는 α 를 구해서 B 에서 구한 식에 대입하여 최종 w 를 구한다. 여기서 α 는 α_i 를 요소로 하는 column vector 이고 $\mathbf{1}$ 은 모든 요소가 1 인 column vector 이다. $L(\alpha)$ 를 최대로 하는 α 를 구하시오.

[10] (각 5 점) Softmax(σ)를 출력노드의 activation 함수로 쓰는 신경망의 loss 는 cross entropy 를 사용한다. 그리고 은닉층 노드의 activation 함수는 사용하지 않는다. 신경망의 출력 노드는 2 개이며, 은닉층은 하나만 있으며, 3 개의 노드를 가지고 있다. 입력층 노드는 2 개이다. 2 개의 입력값을 x_1, x_2 , 출력층의 각 노드의 출력을 o_1, o_2 , 은닉층 각 노드의 출력을 h_1, h_2, h_3 라고 표시한다. 그러면 이 신경망의 matrix-vector form 은 아래와 같다.

$$\begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix} = \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \\ v_{31} & v_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\begin{bmatrix} o_1 \\ o_2 \end{bmatrix} = \begin{bmatrix} \sigma & 0 \\ 0 & \sigma \end{bmatrix} \circ \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix}$$

A. Desired 출력을 t_1, t_2 라고 했을 때 error signal vector $\delta_o = \begin{bmatrix} \delta_{o1} \\ \delta_{o2} \end{bmatrix}$ 를 구하시요.

B. 은닉층 error signal vector $\delta_h = \begin{bmatrix} \delta_{h1} \\ \delta_{h2} \\ \delta_{h3} \end{bmatrix}$ 를 구하시요.

C, D. $W = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix}, V = \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \\ v_{31} & v_{33} \end{bmatrix}$ 라고 했을 때, error backpropagation에 의한 W, V의 update는 아래 식으로 이루어진다.

$$W^{new} = W^{old} + \Delta W, \quad V^{new} = V^{old} + \Delta V.$$

$\Delta W, \Delta V$ 를 learning rate η , $\begin{bmatrix} \delta_{o1} \\ \delta_{o2} \end{bmatrix}, \begin{bmatrix} \delta_{h1} \\ \delta_{h2} \\ \delta_{h3} \end{bmatrix}, \begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix}, \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ 를 이용하여 구하시요. 단, 연산시 dimension이 맞도록 필요시 column vector를 적절하게 row vector로 바꾸어 사용하시요.

[11] (5 점) n 개의 2-차원 벡터 샘플 집합 $\{x_1, \dots, x_n\}$ 의 평균 벡터는 $m = \frac{1}{n} \sum_{k=1}^n x_k$ 로 주어진다. $x_k - m$ 벡터를 크기가 1 인 벡터 e 가 span 하는 공간에 projection 하면 크기가 얼마가 되는가?

[12] (5 점) 2 차원 공간의 데이터로부터 구한 평균 m 이 $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$ 라고 하자. 데이터의 Scatter matrix S 의 orthonormal eigenvector 가 $\begin{pmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \end{pmatrix}, \begin{pmatrix} 1/\sqrt{5} \\ -2/\sqrt{5} \end{pmatrix}$ 이고 각각의 eigenvalue 가 2, 1 일 때, 벡터 $\begin{pmatrix} 4 \\ 3 \end{pmatrix}$ 을 1 차원 값으로 projection 하였을 때 projection error 가 가장 적은 1-st principal component 를 구하시요.

[13] (5 점) Simple linear regression model 이 아래와 같다.

$$Y = \theta_0 + \theta_1 X + \epsilon$$

Correlation coefficient 는 $\rho(Y, X) = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{y_i - \bar{y}}{\sigma_y} \right) \left(\frac{x_i - \bar{x}}{\sigma_x} \right)$ 으로 주어진다. 이때 100 개의 pair data $\{(x_i, y_i) | i = 1, \dots, 100\}$ 로 regression 하였더니 $\theta_0 = 3, \theta_1 = -5$ 로 추정되었다. $\rho(Y, X)$ 이 부호는?

[14] (5 점) 일반적인 regression 문제는 아래와 같이 주어진다.

$$\mathbf{y} = \Phi \theta + \epsilon,$$

이를 Least Squares Estimation을 하면 아래 최적화 문제를 풀어야 한다.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \|\epsilon\|^2 = \|\mathbf{y} - \Phi \theta\|^2 \cong S(\theta)$$

$\hat{\theta}$ 의 closed form solution을 유도하시오.

[15] (5 점) Regression model $y_i = \theta_0 + \theta_1 x_i + \epsilon_i, i = 1, \dots, n$ 의 fitting 정확도를 측정하는 R^2 을 구하는 식을 쓰시오.

[16] (5 점) Elastic Regression 에서 목적함수에 추가되는 regularization 항은 무엇인가?

[17] (각 5점) p 차원의 독립 랜덤 변수 벡터의 관측 값 n개를 column vector로 하는 행렬을 \mathbf{X} 라고 하면 $\mathbf{X} = [x_1 \ x_2 \ \dots \ x_n]$ 과 같다. 평균을 0이 되도록 조정된 값이다.

- A. Data $\{x_1, x_2, \dots, x_n\}$ 의 Scatter matrix S 를 \mathbf{X} 를 이용하여 표현하면?
B. $p \times n$ 차원의 \mathbf{X} 를 $q \times n$ 차원의 \mathbf{Z} 벡터로 차원을 축소하고자 한다. 이를 위해 $\mathbf{Z} = \bar{\mathbf{U}}^T \mathbf{X}$ 를 이용하여 구한다. 다음 중 틀린 내용은?

(1) $\bar{\mathbf{U}}$ 의 column들은 S 의 eigenvalue가 큰 순서대로 q 개에 해당하는 orthonormal eigenvector 들이다.

(1) $\hat{\theta} = (\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{Z}\mathbf{y}$ 이 된다.

(2) $\mathbf{y} = \mathbf{X}^T \theta + \epsilon$ 대신에 $\mathbf{y} = \mathbf{Z}^T \boldsymbol{\theta} + \epsilon$ 를 이용하여 regression을 하면 underfitting 문제를 완화할 수 있다.

(3) $\mathbf{y} = \mathbf{X}^T \theta + \epsilon$ 대신에 $\mathbf{y} = \mathbf{Z}^T \boldsymbol{\theta} + \epsilon$ 를 이용하여 regression을 하면 계산량이 줄어든다.

(4) 틀린 내용 없음

[18] (5 점) Gaussian Process Regression 에서는 $y = f(\mathbf{x}) + \epsilon$ 에서 $f(\mathbf{x})$ 의 분포를 $f(\mathbf{x}) \sim N(\boldsymbol{\mu}_f(\mathbf{x}), \mathbf{k}(\mathbf{x}, \mathbf{x}'))$ 로 가정을 하고 $\boldsymbol{\mu}_f(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$, $\mathbf{k}(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - \boldsymbol{\mu}_f(\mathbf{x}))(f(\mathbf{x}') - \boldsymbol{\mu}_f(\mathbf{x}'))]$ 를 구하는 것이 목적이다. 다음 중 틀린 내용은?

- (1) $f(\mathbf{x})$ 를 모르므로 \mathbf{x} 값 만을 이용하여 $\mathbf{k}(\mathbf{x}, \mathbf{x}')$ 를 구하는 함수를 kernel 이라고 한다.
- (2) Kernel 함수를 $k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2\lambda}(\mathbf{x} - \mathbf{x}')^T (\mathbf{x} - \mathbf{x}')\right)$ 로 선택했을 때, $f(\mathbf{x})$ 와 $f(\mathbf{x}')$ 간의 연관성은 \mathbf{x} 와 \mathbf{x}' 의 거리가 클 수록 낮아지는데, 낮아지는 정도는 λ 값이 작을수록 더 천천히 낮아진다.
- (3) σ_f^2 값이 크면 \mathbf{x} 의 변동에 $f(\mathbf{x})$ 가 크게 변동한다는 것을 의미한다.
- (4) 최적 λ, σ_f^2 는 $p(\mathbf{y}|\mathbf{X})$ 를 최대로 하도록 구한다.
- (5) 틀린 내용 없음

[19] (5점) Kalman filter에서 추정해야 할 상태 값은 $x_k^a = E[x_k | z_k, \dots, z_1]$ 이다. 여기서 $x_k^a = E[x_k | z_{k-1}, \dots, z_1] + E[x_k | z_k] = x_k^f + K_k(z_k - H_k x_k^f)$ 로 구한다. K_k 를 Kalman Gain이라고 한다. 다음 중 틀린 내용은?

- (1) x_k^f 는 $x_k^f = A_{k-1}x_{k-1}^a + B_{k-1}u_{k-1}$ 식으로부터 구한다.
- (2) K_k 는 estimation error covariance matrix $P_k = E[(x_k - x_k^a)(x_k - x_k^a)^T]$ 의 trace 를 최소로 하도록 구한다.
- (3) $z_k - H_k x_k^f$ 값의 의미는 새로 관측된 출력 z_k 와 $k - 1$ 까지의 관측데이터부터 prediction된 출력 \hat{z}_k 간의 오차이다.
- (4) Kalman filter는 visual tracking 문제에도 적용할 수 있다.
- (5) 틀린 내용 없음

[20] (5점) 병원에서 암표지자 검사를 받아서 표지자 값이 $x = 3.2$ 가 나왔다. 이 값으로 posteriori 확률을 계산하였더니 암일 확률이 0.1이다. 정상인을 암환자로 판정할 시 risk를 1으로 하고 암환자를 정상인으로 판단시 risk 는 10로 준다. 정확히 판정하는 경우 risk는 0으로 한다. Expected risk에 기반하여 암 환자 여부를 판단하시오.

[21] (5 점) 동전을 100 번 던져서 head 가 40 번 나오고 tail 이 60 번 나왔다. Head 가 나올 확률 μ_1 과 tail 이 나올 확률 μ_2 를 Bayesian learning 으로 추정하려고 한다.

μ_i 는 Dirichlet 분포를 따른다. Prior 분포 $p(\mu_1, \mu_2 | \alpha_1, \alpha_2)$ 에서 hyper-parameter 가 $\alpha_1 = 10, \alpha_2 = 20$ 로 주어 졌을 때, μ_1, μ_2 를 추정하시오.

[22] (5 점) 관측된 10 개의 word w_i 를 MCMC 로 clustering 하려고 한다. Vocabulary 는 6 개이고, cluster 수는 2 개라 가정한다. Gibbs sampling 을 위해 $p(z_i | z_{-i})$ 를 산출하여 sampling 하여 아래 표와 같이 cluster label 을 assign 하였다. w_i 와 z_i 는 multinomial 분포를 갖는다. w_i 와 z_i 의 분포의 파라미터 ϕ, θ 를 inference 하기위해 아래 표 결과로부터

$$h_\theta(k) = \sum_{i=1}^{10} \delta(z_i - k), \quad h_\phi(k, v) = \sum_{i=1}^{10} \delta(w_i - v) \delta(z_i - k) \text{ 를 산출해야 한다.}$$

$h_\theta(2), h_\phi(1,3), h_\phi(2,3)$ 를 구하시오.

| | | | | | | | | | |
|-------|---|---|---|---|---|---|---|---|---|
| i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| w_i | 1 | 3 | 2 | 3 | 3 | 5 | 4 | 1 | 6 |
| z_i | 1 | 2 | 2 | 2 | 1 | 1 | 2 | 1 | 2 |