Quiz 1: 2,3 강 퀴즈

문제 1

정답

총 1,00 점에서 1,00 점 할당

질문 편집

If v_i and λ_i are eigenvectors and eigenvalues of a matrix A, the nullity of $(A - \lambda_i \mathbf{I})$ is more than or equal to one.

하나를 선택하세요.

참 ✓

거짓

정답 : '참'

문제 2

정답

총 1,00 점에서 1,00 점 할당

질문 편집

$\{(2,1,0),(1,0,1),(0,0,3),(1,1,-1)\}$ is a generating set of 3-dimensional real space.

하나를 선택하세요.

참 ✓

거짓

$\{(2,1,0),(1,0,1),(0,0,3)\}$ can be a basis of 3-dimensional real space.

정답 : '참'

문제 3

정답

총 1,00 점에서 1,00 점 할당

질문 편집

$\{x+1, 2x+1, x^2+2\}$ is a linearly independent set.

하나를 선택하세요.

참

거짓

정답 : '참'

문제 **4**

정답

총 1,00 점에서 1,00 점 할당

질문 편집

If eigenvalues, λ_i , $i = 1, \ldots, n$, of a matrix $\boldsymbol{A}$ are distinctive, $\boldsymbol{A}$ is always diagonalizable.

하나를 선택하세요.

- 참 ✓
- 거짓

정답 : '참'

문제 5

정답

총 1,00 점에서 1,00 점 할당

질문 편집

A solution set of ₩(Ax=b, b ₩neq 0, can be a vector space.

하나를 선택하세요.

참

거짓

A solution set is $\{x_p\} + N(A)$, where x_p is a particular solution. Thus It does not zero vector (origin, + identity) and so can not be vector space.

정답 : '참'

문제 6

정답

총 1,00 점에서 1,00 점 할당

질문 편집

A set of continuous functions is a subspace of a set of analytic functions.

하나를 선택하세요.

참 ✓

거짓

A set of analytic functions is a subspace of a set of continuous functions.

정답 : '참'

문제 7

정답

총 1,00 점에서 1,00 점 할당

질문 편집

We can find elementary row operation matrix E_i such that $A^{-1} = E_n \cdots E_2 E_1 \mathrm{I}$

하나를 선택하세요.

참 ✓

거짓

If A is not invertible, the statement is not true.

정답 : '참'

문제 8

정답

총 1.00 점에서 1.00 점 할당

질문 편집

For $X^T := [x_1 \quad x_2]$, where $x_1 = \begin{bmatrix} 1 \\ -2 \end{bmatrix}, x_2 = \begin{bmatrix} -4 \\ 3 \end{bmatrix}$,
Calculate $X^T X$. Then

$$X^T X = \begin{bmatrix} x_1^T \\ x_2^T \end{bmatrix} [x_1 \quad x_2] = \begin{bmatrix} x_1^T x_1 & x_1^T x_2 \\ x_2^T x_1 & x_2^T x_2 \end{bmatrix}$$

하나를 선택하세요.

참

거짓 ✓

$$X^T X = [x_1 \quad x_2] \begin{bmatrix} x_1^T \\ x_2^T \end{bmatrix} = x_1 x_1^T + x_2 x_2^T$$

정답 : '거짓'

문제 9

정답

총 1,00 점에서 1,00 점 할당

질문 편집

If the Hessian matrix of $f = \frac{1}{2}(x^2 + 2xy + y^2)$ which is a function of a vector (x, y), the Hessian matrix is positive definite and the minimum point is unique.

하나를 선택하세요.

참

거짓 ✓

$f = \frac{1}{2}(x^2 + 2xy + y^2),\ H = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix},\ v^T H v = (v_1 + v_2)^2 \geq 0$
$\rightarrow H \geq 0 \rightarrow f = (x + y)^2$ has minimum at $x = -y$

정답 : '거짓'

문제 10

정답

총 1,00 점에서 1,00 점 할당

질문 편집

If a non-invertible linear equation $Ax = b,\ \rho(A) < n,\ x \in R^n$ has many solutions, the minimal solution with minimum norm $\|x\|$ can be expressed by $x = A^{\dagger}b$, where $A^{\dagger}$: pseudo inverse.

하나를 선택하세요.

참 ✓

거짓

정답 : '참'

문제 11

정답

총 1,00 점에서 1,00 점 할당

질문 편집

A null space of $A = \begin{bmatrix} 1 & -1 & 3 \\ 2 & -2 & 6 \end{bmatrix}$ is one-dimensional space.

하나를 선택하세요.

참

거짓 ✓

$$N(A) = \left\{ \begin{bmatrix} x \\ y \\ z \end{bmatrix} | x - y + 3z = 0 \right\} = \left\{ s \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} -3 \\ 0 \\ 1 \end{bmatrix} | s, t \in R \right\}, \qquad \text{dim.} = 2, \text{basis} = \left\{ \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -3 \\ 0 \\ 1 \end{bmatrix} \right\}$$

정답 : '거짓'

Quiz 2: 4,5 강 퀴즈

문제 1

정답

총 1,00 점에서 1,00 점 할당

질문 편집

Conditional Entropy $H(X|Y)$ 는 Y 와 연관이 있는 X 의 평균 정보량을 의미한다.

하나를 선택하세요.

참

거짓 ✓

Conditional Entropy $H(X|Y)$ 는 Y 와 연관이 없는 X 의 평균 정보량을 의미한다.
Y 와 연관이 있는 X 의 평균 정보량은 mutual information $I(X,Y)$ 이다.
정답 : '거짓'

문제 2

정답

총 1,00 점에서 1,00 점 할당

질문 편집

KL-Divergence $D_{P(X|Y=0)||P(X|Y=1)}$ 가 최소가 될 때 X 와 Y 의 mutual information I(X,Y) 이 최대가 된다.

하나를 선택하세요.

참

거짓 ✓

KL-Divergence $D_{P(X|Y=0)||P(X|Y=1)}$ 가 최소값 0을 가지면 X 와 Y 가 서로 independent 가 되고 mutual information I(X,Y) 가 0이 되어 최소가 된다.

정답 : '거짓'

문제 3

정답

총 1,00 점에서 1,00 점 할당

질문 편집

Let $L(\theta, \lambda, \nu)$ be a Lagrangian for a convex problem. If θ, λ, ν satisfy KKT for the convex problem, they are optimal.

하나를 선택하세요.

참 ✓

거짓

정답 : '참'

문제 4

정답

총 1,00 점에서 1,00 점 할당

질문 편집

The linear equation $\begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \theta^* \\ v^* \end{bmatrix} = \begin{bmatrix} -q \\ b \end{bmatrix}$ has a unique solution if and only if $P + A^T A \succ 0$.

하나를 선택하세요.

- 참 ✓
- 거짓

정답 : '참'

문제 5

정답

총 1,00 점에서 1,00 점 할당

질문 편집

$C_{p||q}(x;W) = -\sum_{x\in D}\sum_k p_k(x)\ \log q_k(x;W)$ 는 Cross entropy loss 이며 D는 학습데이터의 집합, $q_k(x;W)$는 $x \in D$ 를 입력으로 하는 신경망의 k번째 출력 노드 값, $p_k(x)$ 는 k번째 출력 노드 값의 desired 값, W는 신경망의 연결 가중치이다. 이 loss 는 multi-label classification 문제에 적용하게 된다.

하나를 선택하세요.

참

거짓 ✓

이 loss 는 one-hot classification 문제에 적용할 수 있음.

정답 : '거짓'

문제 6

정답

총 1,00 점에서 1,00 점 할당

질문 편집

랜덤 변수 X에 대한 entropy $H(X)$ 가 최대가 되는 경우는 X의 어느 하나의 instance만 발생하고 나머지는 발생할 확률이 0이 될 때이다.

하나를 선택하세요.

- 참
- 거짓 ✓

랜덤 변수 X에 대한 entropy $H(X)$ 가 최대가 되는 경우는 X의 모든 instance가 발생할 확률이 같을 때이다.

정답 : '거짓'

문제 7

정답

총 1.00 점에서 1.00 점 할당

질문 편집

다음이 convex optimization 문제가 주어진다.

minimize $\theta^T\theta, \theta \in R^n$

subject to $A\theta \leq b,\ b \in R^p,\ p \leq n,\ A$ has full rank.

이 문제의 Lagrangian 은 $L(\theta, v) = \theta^T\theta - v^T(A\theta - b)$ 로 주어지며 v 의 모든 element는 양수이다.

하나를 선택하세요.

참

거짓 ✓

Lagrangian 은 $L(\theta, v) = \theta^T\theta + v^T(A\theta - b)$ 로 주어짐.

정답 : '거짓'

문제 8

정답

총 1,00 점에서 1,00 점 할당

질문 편집

2nd-order conditions: for twice differentiable f with convex domain

f is convex if and only if

$$\nabla^2 f(x) \geq 0 \quad \text{for all } x \in \mathbf{dom}\, f$$

If $\nabla^2 f(x) > 0$ for all $x \in \mathbf{dom}\, f$, then f is strictly convex

하나를 선택하세요.

참 ✓

거짓

정답 : '참'

문제 9

정답

총 1,00 점에서 1,00 점 할당

질문 편집

Entropy is defined by a measure of the amount of information conveyed by a message.

하나를 선택하세요.

참

거짓 ✓

Entropy is defined by a measure of the average (expected) amount of information conveyed per message.

정답 : '거짓'

Quiz 3: 6,7 강 퀴즈

(o) Linear classifier $g(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$ 가 주어 졌을 때, $g(\boldsymbol{x}) = \pm 1$이 되는 $\boldsymbol{x}$ 를 support vector 라고 한다.

(x) Support vector machine 의 목적은 모든 학습데이터 $\{\boldsymbol{x}_i \mid i = 1, \dots, n\}$ 가 $|g(\boldsymbol{x}_i)| \geq 1$ 인 영역에 존재하고, $|g(\boldsymbol{x})| < 1$인 $\boldsymbol{x}$ 의 영역이 최대가 되도록 Linear classifier $g(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$ 의 법선벡터 $\boldsymbol{w}$ 와 절편 b 를 구하는 것이며, 이는 $|g(\boldsymbol{x}_i)| \geq 1$ 조건에서 $\|\boldsymbol{w}\|$ 를 최대로 하는 것이다.

(o) $\|\boldsymbol{w}\|$ 를 최소로 하는 것이 됨.

(o) Support vector machine 에서 학습데이터 $\{\boldsymbol{x}_i \mid i = 1, \ldots, n\}$ 를 사용하여 구한 Linear classifier $g(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$ 의 법선벡터 $\boldsymbol{w}$ 는 $\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i\, y_i \boldsymbol{x}_i$ 로 구해지며, α_i 는 $\sum_{i=1}^{n} \alpha_i y_i = 0$을 만족하는 값 중에서 $L_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i{}^T \mathbf{x}_j$ 를 최대로 하는 양수 값이 된다.

(x) Support vector machine 에서 학습데이터 $\{\boldsymbol{x}_i \mid i = 1, \ldots, n\}$ 를 사용하여 구한 Linear classifier $g(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$ 의 법선벡터 $\boldsymbol{w}$ 는 $\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i\, y_i \boldsymbol{x}_i$ 로 구해지며, $g(\mathbf{x}_i) = \pm 1$을 만족하는 $\boldsymbol{x}_i$에 곱해지는 α_i는 complement slackness에 의해 모두 0이 된다.

(o) $g(\mathbf{x}_i) = \pm 1$을 만족하는 $\boldsymbol{x}_i$에 곱해지는 α_i만 0 아니고 나머지는 모두 0이 됨.

(o) Support vector machine 에서 학습데이터 $\{\mathbf{x}_i \mid i = 1, \dots, n\}$ 를 사용하여 α_i 를 구할 때 $L_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i{}^T \mathbf{x}_j$ 에서 $\mathbf{x}_i{}^T \mathbf{x}_j$ 를 kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$ 를 사용하면 $\mathbf{x}$ 공간에서의 class 간의 경계면을 임의의 곡면으로 학습할 수 있게 되어 선형이 아닌 복잡한 경계면을 갖는 classification 문제를 학습할 수 있다.

(x) Soft margin 을 사용하는 support vector machine 에서는 학습데이터 $\{\boldsymbol{x}_i \mid i = 1, \dots, n\}$ 중에서 $|g(\boldsymbol{x})| < 1$ 의 영역에 학습데이터가 존재하는 것을 허용하도록 하는 것이다. 이 경우는 Linear classifier $g(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + b$ 의 법선벡터 $\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i$ 의 α_i는 $0 \le \alpha_i \le C$ 로 제한이 되고, $|g(\boldsymbol{x})| < 1$ 의 영역에 존재하는 $\boldsymbol{x}_i$ 에 곱해지는 α_i의 값은 $0 < \alpha_i < C$ 가 된다.

(o) $|g(\boldsymbol{x})| < 1$ 의 영역에 존재하는 $\boldsymbol{x}_i$ 에 곱해지는 α_i의 값은 $\alpha_i = C$ 가 됨.

(o) 신경망의 k 번째 출력 노드의 값이 $a_k = w_k^T h = \sum_j w_{kj} h_j$ 로 주어 지고 teacher 값은 t_k 로 주어지는 regression 문제에서 Loss 함수는 $E(w) = {}^1/_2 \sum_k (a_k - t_k)^2$ 로 주어진다. 학습시 w_{kj} 업데이트 식은 $\Delta w_{kj} = -\eta \frac{\partial E_d}{\partial w_{kj}} = -\eta \frac{\partial E_d}{\partial a_k} \frac{\partial a_k}{\partial w_{kj}} = -\eta \frac{\partial E_d}{\partial a_k} h_j = \eta (t_k - a_k)\, h_j$ 가 된다.

(x) Forward pass 가 vector form 으로 $b = Vx \rightarrow h = r(b) \rightarrow a = Wh \rightarrow o = \sigma(a)$ 와 같이 나타내는 신경망에서, W 와 V의 업데이트 식은 $W^{new} = W^{old} + \eta(t - o)h$, $V^{new} = V^{old} + \eta Diag(r'(b) W (t - o)x$로 주어진다.

(o) $W^{new} = W^{old} + \eta(t - o)h^T$, $V^{new} = V^{old} + \eta Diag(r'(b) W^T (t - o)h^T$로 주어짐.

(o) Convolution neural network 의 은닉층의 channel 수는 연결된 filter의 수와 일치한다.

(x) Dropout은 연결가중치 학습시 랜덤하게 선택된 연결가중치를 학습에서 배제시키는 방법으로 업데이트 가중치 수를 줄여서 학습 연산량을 줄이기 위한 것이 주 목적이다.

(o) 일반화 성능을 높이기 위한 것이 주 목적임.

(o) Convolution neural network 은 하위층일 수록 edge, blob, texture 등 일반적인 특징을 학습하고 상위층으로 갈 수록 feature binding 이 이루어져서 물체단위의 특징을 학습하는 것으로 알려져 있다.

(x) Pretrained 된 convolution neural network 을 활용하여 데이터 량이 적은 새로운 task를 fine tuning 할 때, 상위층의 학습율을 작게 하고 상위층의 학습율을 상대적으로 크게 하면 모든 층의 학습율을 같게 하는 것보다 성능이 더 좋은 것으로 알려져 있다.

(o) 하위층의 학습율을 작게 하고 상위층의 학습율을 더 크게 하는 것이 더 좋은 성능을 보임.

Quiz 4: 8,9 강 퀴즈

(o) Principal Components Analysis (PCA) seeks the projection that minimizes the sum of squared errors between original data and their projections to a lower dimensional space.

(x) Linear Discriminant Analysis (LDA) seeks the projection that minimizes between-class distance and maximizes within-class distance after projections.

(o) Linear Discriminant Analysis (LDA) seeks the projection that maximizes between-class distance and minimizes within-class distance after projections.

(o) 2차원 공간에서 vector $\boldsymbol{m}$ 을 지나는 직선 중 unit vector $\boldsymbol{e}$ 방향과 같은 직선위의 vector $\boldsymbol{x}$ 는

$\boldsymbol{x} = \boldsymbol{m} + a\boldsymbol{e}$ 와 같이 표현된다.

(x) 3차원 공간에서 vector $\boldsymbol{m}$ 을 지나는 평면 중 unit vector $\boldsymbol{e_1}$ 과 $\boldsymbol{e_2}$ 가 span 하는 평면 W 가 있다.

3차원 공간의 임의 점 vector $\boldsymbol{x}$ 를 평면 W에 projection 한다. 이 projection 한 vector를 $\boldsymbol{m}$ 을 원점으로 하고

$\boldsymbol{e_1}$ 과 $\boldsymbol{e_2}$를 coordinates 로 하여 representation 하면 $[\boldsymbol{e_1^T x} \quad \boldsymbol{e_2^T x}]$ 가 된다.

(o) $[\boldsymbol{e_1^T}(\boldsymbol{x} - \boldsymbol{m}) \quad e_2^T(\boldsymbol{x} - \boldsymbol{m})]$이 됨.

(o) $S = \sum_{k=1}^{n}(x_k - m)(x_k - m)^T$ 는 scatter matrix 라 칭하며, 주어진 데이터 $\{x_k | k = 1, \dots, n\}$의 기하 학적 분포를 나타낸다. m을 중심으로 각 vector x_k의 위치하는 공간을 나타내는 matrix이며, S의 eigenvalue 가 가장 큰 eigenvector 방향으로 데이터 샘플이 많이 퍼져 있다고 할 수 있다.

(x) PCA 에서 projection 시 가장 적은 오차를 나타내는 coordinate $\boldsymbol{e}$ 를 찾기 위해서는 scatter matrix 를 S 라 했을 때, $-\boldsymbol{e}^T S \boldsymbol{e}$ 를 최대로 하는 $\boldsymbol{e}$ 를 구하면 되고 이는 S의 최대 eigenvalue에 해당하는 unit eigenvector 를 구하면 된다.

(o) $\boldsymbol{e}^T S \boldsymbol{e}$ 를 최대로 해야함.

(o) Scatter matrix를 S이 eigenvalue를 큰 순서 대로 indexing 하여 $\{\lambda_i | \, i = 1, \dots, 100\}$ 라 했을 때, 100차원을 10차원공간으로 PCA로 projection 하면 손실 오차의 총 합은 $\sum_{i=11}^{100} \lambda_i$ 가 된다.

(o) i-th class의 scatter matrix $S_i = \sum_{x \in C_i}(x - m_i)(x - m_i)^T$ 를 $\boldsymbol{w}$ 라는 vector 로 나타내는 축으로 projection 하면 $\boldsymbol{w}^T S_i \boldsymbol{w}$ 가 된다.

(x) i-th class의 data 의 중심점 m_i 를 $\boldsymbol{w}$ 라는 vector 로 나타내는 축으로 projection 하면 $\boldsymbol{w}^T m_i$ 가 된다.
따라서 두 class 의 within-class distance 를 구하면 $\boldsymbol{w}^T(m_1 - m_2)(m_1 - m_2)^T \boldsymbol{w}$ 가 된다.

(o) between-class distance 임.

(o) LDA 기법을 이용하여 두 class 를 가장 잘 구분하도록 하는 projection 축 $\mathbf{w}$ 는 $S_W^{-1}(m_1 - m_2)$ 의 방향을 갖는다.

(o) i-th class의 scatter matrix $S_i = \sum_{x \in C_i}(x - m_i)(x - m_i)^T$ 를 $\boldsymbol{w}$ 라는 vector 로 나타내는 축으로 projection 하면 $\boldsymbol{w}^T S_i \boldsymbol{w}$ 가 된다.

(x) i-th class의 data 의 중심점 m_i 를 $\boldsymbol{w}$ 라는 vector 로 나타내는 축으로 projection 하면 $\boldsymbol{w}^T m_i$ 가 된다.
따라서 두 class 의 within-class distance 를 구하면 $\boldsymbol{w}^T(m_1 - m_2)(m_1 - m_2)^T \boldsymbol{w}$ 가 된다.

(o) between-class distance 임.

(o) LDA 기법을 이용하여 두 class 를 가장 잘 구분하도록 하는 projection 축 $\mathbf{w}$ 는 $S_W^{-1}(m_1 - m_2)$ 의 방향을 갖는다.

(o) Random variable Y, X 에 대해 $Y = \theta_0 + \theta_1 X + \epsilon$, $\epsilon \sim N(0, \sigma^2)$ 의 관계로 가정된 회기 모델에서
X와 Y의 covariance 의 부호와 θ_1의 부호는 일치한다.

(x) Regression 문제는 $y_i = \theta_0 + \theta_1 x_i + \epsilon_i,\ i = 1, \cdots, n, \quad \epsilon_i \sim N(0, \sigma^2)$로 가정된 모델에서
주어진 데이터 $\{(x_i, y_i) |\ i = 1, \cdots, n\}$ 로 부터 매개변수 $\theta_0,\ \theta_1$을 추정하는 것이다.
참값과 추정오차가 가장 적은 방법은

$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(y_i - \theta_0 - \theta_1 x_i)^2}{2\sigma^2})$ 또는 $\sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$ 를 최소로 하는 $\theta_0,\ \theta_1$을 구하면 된다.

(o) $\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(y_i - \theta_0 - \theta_1 x_i)^2}{2\sigma^2})$ 는 최대로 하는 $\theta_0,\ \theta_1$을 구해야 함.

(o) Multivariate linear regression model $Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \cdots + \theta_p X_p + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$ 에서

추정된 매개변수를 $\{\hat{\theta}_i \mid i = 1, \ldots, n\}$ 라 하면 Y의 추정 값은 $\hat{Y} = \hat{\theta}_0 + \hat{\theta}_1 X_1 + \hat{\theta}_2 X_2 + \cdots + \hat{\theta}_p X_p$ 이 된다.

이때 Y 와 $\hat{Y}$ 의 correlation coefficient 가 1에 가까울 수록 추정 값이 정확하다고 할 수 있다.

(x) Multivariate linear regression model 을 vector form 으로 나타내면 $y_i = \boldsymbol{x}_i^T \boldsymbol{\theta} + \epsilon_i, \; i = 1, \cdots, n$ 로 나타낼 수 있다.

여기서 $\boldsymbol{\theta}^T = [\theta_0 \;\; \theta_1 \;\; \cdots \;\; \theta_p]$, $\boldsymbol{x}_i^T = [x_{i0} \;\; x_{i1} \;\; \cdots \;\; x_{ip}]$ 로 표현된다.

이때 $\boldsymbol{y} = [y_1 \;\; y_2 \;\; \cdots \;\; y_n]^T$, $\boldsymbol{\epsilon} = [\epsilon_1 \;\; \epsilon_2 \;\; \cdots \;\; \epsilon_n]^T$ 를 정의하고, $\boldsymbol{X} = [\boldsymbol{x}_1 \, \boldsymbol{x}_2 \;\; \cdots \;\; \boldsymbol{x}_n]^T$ 로 표시하면

Regression model은 $\boldsymbol{y} = \boldsymbol{X\theta} + \boldsymbol{\epsilon}$ 의 matrix-vector form 으로 표현할 수 있다.

오차를 최소로 하는 매개변수 추정 값 $\hat{\boldsymbol{\theta}}$는 $\hat{\boldsymbol{\theta}} = (\boldsymbol{XX}^T)^{-1}\boldsymbol{Xy}$ 이 된다.

(o) $\hat{\boldsymbol{\theta}} = (\boldsymbol{XX}^T)^{-1}\boldsymbol{Xy}$ 는 연산 차원이 맞지 않음. $\hat{\boldsymbol{\theta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ 이 되어야 함.

Quiz 5: 10, 11 강 퀴즈

(o) Matrix form 으로 표현되는 linear regression model 은 $\boldsymbol{y} = \Phi\theta + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I})$ 로 주어진다.
매개변수 벡터가 p 차원 일 때 RLS 추정치를 $\hat{\theta}$ 라 하자. 그러면 residual vector $\boldsymbol{e} = \boldsymbol{y} - \Phi\hat{\theta}$ 의 variance $E(\boldsymbol{e}^T\boldsymbol{e})$ 는 $(n-p)\sigma^2$이 된다.

(x) ϕ_i 가 p 차원 벡터일 때, $y_i = \phi_i^T\theta + \epsilon_i$ 로 주어지는 regression 모델의 매개변수 θ 를 Recursive Least Squares 로 추정하려고 한다. 행렬 Φ_k를 $\Phi_k = [\phi_1\ \phi_2\ \cdots\ \phi_k]^T$ 로, $\boldsymbol{y}_k$를 $\boldsymbol{y}_k = [y_1\ y_2\ \cdots\ y_k]^T$ 로 정의한다. 이때 $k+1$ 번째 추정치 $\hat{\theta}_{k+1}$은
$\hat{\theta}_{k+1} = (\Phi_k^T\Phi_k + \phi_{k+1}^T\phi_{k+1})^{-1}\Phi_{k+1}^T\boldsymbol{y}_{k+1}$ 로 주어진다.

(o) $\hat{\theta}_{k+1} = (\Phi_k^T\Phi_k + \phi_{k+1}\phi_{k+1}^T)^{-1}\Phi_{k+1}^T\boldsymbol{y}_{k+1}$ 이 되어야 함. $\phi_{k+1}^T\phi_{k+1}$는 스칼라로 차원이 맞지 않음.

(o) Zero mean으로 정규화한 regression vecto를 $\mathbf{x}_i$라 정의하였을 때 **scatter matrix**를 $\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_n]$ 로 정의한다. Nonlinear Iterative Partial Least Squares (NIPALS) algorithm은 임의의 regression vector $\mathbf{x}_j$ 를 $\mathbf{t}$에 할당하고 $\mathbf{u} = \mathbf{Xt}/\|\mathbf{Xt}\|$ 와 $\mathbf{t} = \mathbf{X}^T\mathbf{u}$ 를 번갈아 loop를 수행하여 $\mathbf{t}$가 변화하지 않으면 중지하고 $\mathbf{u}$를 제 1 principle component로 설정한다.
$\mathbf{X}^T \coloneqq \mathbf{X}^T - \mathbf{tu}^T = \mathbf{X}(\mathrm{I} - \mathbf{uu}^T)$로 **scatter matrix**를 변경하고 위 loop를 반복하여 $\|\mathbf{Xt}\|$ 가 충분히 작을 때 까지 순차적으로 후속 principle components를 찾아낸다.
이 $\mathbf{u}$ 들로 구성된 변환 행렬로 $\mathbf{X}$의 차원을 축소하여 partial RLS 알고리즘을 수행한다.

(x) Mean Squared Error $MSE(\hat{\theta}) = E\left((\hat{\theta} - \theta)^2\right)$ 는

$MSE(\hat{\theta}) = \left(E(\hat{\theta}) - \theta\right)^2 + E\left(\hat{\theta} - E(\hat{\theta})\right)^2 = Bias(\hat{\theta}, \theta)^2 + Var(\hat{\theta})$ 로 표현될 수 있다.

이때 $Bias(\hat{\theta}, \theta)^2$ 가 크면 overfitting 이 된 것을 의미하며, $Var(\hat{\theta})$가 크면 underfitting 이 된 것을 의미한다.

(o) $Bias(\hat{\theta}, \theta)^2$ 가 크면 underfitting 이 된 것을 의미하며, $Var(\hat{\theta})$가 크면 overfitting 이 된 것을 의미한다.

(o) Weighted recursive least squares 알고리듬은 $\Phi_k = [\lambda^{k-1}\phi_1 \;\; \lambda^{k-2}\phi_2 \;\; \cdots \;\; \phi_k]^T$, $\lambda < 1$,

로 하여 과거의 observation 의 비중을 줄이고 최근 observation 비중을 높여서

regression 하는 방법이다. 이에 Ridge regression 을 적용하면, $\|\boldsymbol{y}_k - \Phi_k\theta\|^2 + \gamma\|\theta\|_2^2$ 을

최소로 하는 θ 을 recursive form으로 구하는 formula를 유도하면 다음과 같다.

$$\hat{\theta}_{k+1} = \hat{\theta}_k + G_k(\boldsymbol{y}_{k+1} - \phi_{k+1}^T\hat{\theta}_k),$$

$$G_k \cong \frac{\lambda^{-1}P_k\phi_{k+1}}{1+\lambda^{-1}\phi_{k+1}^T P_k\phi_{k+1}}$$

$$P_{k+1} = \lambda^{-1}P_k - \lambda^{-1}G_k\phi_{k+1}^T P_k, \; P_0 = -\gamma\mathbf{I}$$

(x) Lasso regression 은 $\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\|\boldsymbol{y} - \Phi\theta\|^2 + \gamma\|\theta\|_1$ 을 최소로 하는 방법이다.

이때, $\|\theta\|_1 = \sum_{i=1}^{p}|\theta_i|$ 는 미분이 불가능하다. 이를 해결하기 위해 $|\theta_i| = s_i$로 하여

다음과 같이 constraint optimization formulation 으로 변경하여 해를 구한다.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\|\boldsymbol{y} - \Phi\theta\|^2 + \gamma\mathbf{1}^T\boldsymbol{s}$$

subject to $|\theta_i| = s_i, \;\; i = 1, \cdots, n$

(o) subject to $|\theta_i| \le s_i, \;\; i = 1, \cdots, n$ 이어야 함.

(o) $y = f(\mathbf{x}) + \epsilon$, $\epsilon \sim N(0, \sigma^2 \mathbf{I})$ 에서 $f(\mathbf{x})$ 값 추정을 위한 Gaussian process regression 은 $f(\mathbf{x})$ 가 mean function이 $\mu_f(\mathbf{x})$, autocovariance function이 $k(\mathbf{x}, \mathbf{x}') = \mathbb{E}\left[\left(f(\mathbf{x}) - \mu_f(\mathbf{x})\right)\left(f(\mathbf{x}') - \mu_f(\mathbf{x}')\right)\right]$ 인 gaussian process 로 가정을 한다. 이때 $f(\mathbf{x})$가 직접 observation 이 안되므로, 이 $k(\mathbf{x}, \mathbf{x}')$를 observation $\mathbf{x}, \mathbf{x}'$로 부터 구하는 kernel $k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \mathbf{exp}\left(-\frac{1}{2\lambda}(\mathbf{x} - \mathbf{x}')^T (\mathbf{x} - \mathbf{x}')\right)$ 로 대체하여 추정하게 된다.

(x) $y = f(\mathbf{x}) + \epsilon$, $\epsilon \sim N(0, \sigma^2 \mathbf{I})$ 에서 $f(\mathbf{x})$ 값 추정을 위한 Gaussian process regression에 사용되는 kernel $k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \mathbf{exp}\left(-\frac{1}{2\lambda}(\mathbf{x} - \mathbf{x}')^T (\mathbf{x} - \mathbf{x}')\right)$ 에서 σ_f^2 은 $\mathbf{x}$와 $\mathbf{x}'$ 간의 correlation 정도를 나타내는 hyperparameter 이고 λ는 $f(\mathbf{x})$의 변동량과 $\mathbf{x}$ 변동량의 관계를 나타낸다.

(o) σ_f^2 은$f(\mathbf{x})$의 변동량과 $\mathbf{x}$ 변동량의 관계를 나타내고, λ는 $\mathbf{x}$와 $\mathbf{x}'$ 간의 correlation 정도를 나타낸다.

.

(o) Kalman filter는 Stochastic time-variant linear system에서 system matrix를 알고 있을 때,
측정가능한 system 입, 출력으로 부터 system state 를 추정하는 것이다. 이때,
model uncertainty와 measurement noise 의 영향을 최소로 하는 것이 목표이다.

(x) $y = f(\mathbf{x}) + \epsilon$, $\epsilon \sim N(0, \sigma^2 \mathbf{I})$ 에서 $f(\mathbf{x})$ 값 추정을 위한 Gaussian process regression에 사용되는
kernel $k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \mathbf{exp}\left(-\frac{1}{2\lambda}(\mathbf{x} - \mathbf{x}')^T (\mathbf{x} - \mathbf{x}')\right)$ 에서 σ_f^2, λ는 $\log p(\mathbf{X}|\mathbf{y})$ 를 최대로 하는 값으로
결정할 수 있다.

(o) σ_f^2, λ는 $\log p(\mathbf{y}|\mathbf{X})$ 를 최대로 하는 값으로 결정함.

.

(o) Kalman filter에서 $k-1$ step 까지 측정된 출력 집합 Z_{k-1}로 부터 상태 $\mathbf{x}_k$ 의 예측치

$\mathbf{x}_k^f = E[\mathbf{x}_k|Z_{k-1}]$ 로 부터 새로 측정된 출력 z_k 로 부터 추가 보정한 $\mathbf{x}_k$ 의 추정치는

$\mathbf{x}_k^a = \mathbf{x}_k^f + E[\mathbf{x}_k|z_k]$ 로 구해진다. 이때 $E[\mathbf{x}_k|z_k]$ 는 출력의 추정오차에 비례하여

$E[\mathbf{x}_k|z_k] = K_k(z_k - \hat{z}_k) = K_k\left(z_k - H_k\mathbf{x}_k^f\right)$ 로 구해진다. 이때 K_k 를 Kalman Gain이라 부른다.

(x) Kalman Gain K_k 는 $E\left[\left(\mathbf{x}_k - \mathbf{x}_k^f\right)\left((\mathbf{x}_k - \mathbf{x}_k^f\right)^T\right]$의 trace 를 최소로 하는 값으로 구한다.

(o) $E\left[\left(\mathbf{x}_k - \mathbf{x}_k^f\right)\left((\mathbf{x}_k - \mathbf{x}_k^f\right)^T\right]$는 K_k 를 포함하지 않으며, $E\left[(\mathbf{x}_k - \mathbf{x}_k^a)\left((\mathbf{x}_k - \mathbf{x}_k^a\right)^T\right]$를 최소로 하는 K_k 를 구함.

.

Quiz 6: 12, 13 강 퀴즈

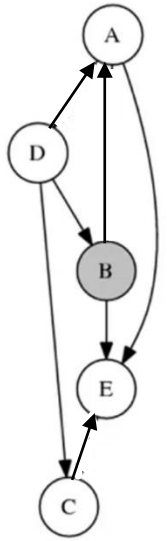
Q: 옆 그림에서 D에서 E로 가는 path 중에서 blocked 인 경우를 고르시오..

(1) $D - C - E$

(2) $D - A - E$

(3) $D - B - A - E$

A: (3) (B의 관측에 의해 blocked 로 변환됨)



Q: 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 risk 에 기반하여 암 환자 여부를 판단하시오.

(1) 암환자 (2) 정상인 (3) 판단하지 못함

A: (1)

설명:

정상으로 판단 리스크

risk $R(\alpha_1|x = 3.2) = \lambda_{12} p(\text{암}|x = 3.2) = 10 \times 0.1$

암환자로 판단 리스크

risk $R(\alpha_2|x = 3.2) = \lambda_{21} p(\text{정}|x = 3.2) = 1 \times 0.9$

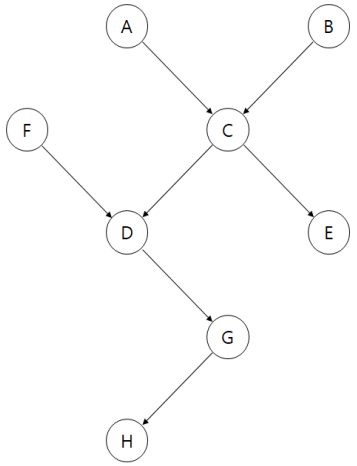
$R(\alpha_2|x = 3.2) < R(\alpha_1|x = 3.2)$ 이므로 암환자로 판정한다.

Q: 우측 그림에서 노느 C를 추론 (inference) 하려고 한다. 즉, 다른 노드의 랜덤 변수의 결과값 (outcome)이 관측되었을 때 C의 특정 사건(event) 이 발생할 확률을 구하려고 한다. 이를 위해 반드시 관측이 되지 않아도 되는 노드를 다음 보기 중에서 고르시오.

(1)A (2) E (3) F (4) G

A: (4)

설명: C 와 독립인 노드는 관측값의 결과에 영향을 받지 않는다. 따라서 Markov Blanket의 범위에 있는 노드만 조건(condition)으로 주어 지면 된다. 즉, C 의 parent, child, child의 parent 만 필요하다. 따라서, A, B, D, E, F, 여기에 포함되지 않는 노드는 (4) G이다.



Q: 병원에 암진단을 받으러 온 사람은 N명이다. 그중 n명이 암환자로 판명이 난다. 암진단은 피검사를 하여 암표지자 값을 보고 판단을 한다. 이 암표지자를 확률변수 x 로 했을 때, 정상인과 암환자 모두 다음의 매개변수 θ 로 표현되는 확률분포는 $p\,(x|\,\theta) = \theta e^{-\theta x}$, $for\ x > 0\ and\ \theta > 0$의 형태를 가지고 각 샘플은 i.i.d. 라고 가정한다. 정상인의 암표지자 값을 평균하면 0.05이 되고 암환자의 암표지자를 평균하면 0.2이 된다. 암환자의 분포를 가장 잘 나타내는 θ 를 MLE 방법으로 추정하면 무엇인지 고르시오.

(1) 100 (2) 10 (3) 5 (4) 20

A: (3)

$$p(D_i|\theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i}$$

$$l(\theta) = \log p(D_i|\theta) = n \log \theta - \theta \sum_{i=1}^{n} x_i$$

$$\frac{d}{d\theta} l(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0, \quad \rightarrow \quad \hat{\theta}_{MLE} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{0.2} = 5$$

Q: 관측된 특징에 대해 Class 를 잘 맞추었을 때 risk를 0, 못 맞추었을 때 risk 를 1로 하면 관측된 특징 x 에 대해 i class 하고 판단했을 때 conditional risk $R(\alpha_i|x)$ 는 보기 중 어느 값인가?

(1) $p(x|\omega_i)$ (2) $1-p(x|\omega_i)$ (3) $p(\omega_i|x)$ (4) $1-p(\omega_i|x)$

A: (4)

설명: $R(\alpha_i|x)=\sum_{j=1}^{c}\lambda(\alpha_i|\omega_j)p(\omega_j|x)=\sum_{j\neq i}^{c}p(\omega_j|x)=1-p(\omega_i|x)$

conditional risk를 최소로 하는 것과 posterior probability 를 최대로 하는 것과 같음.

Q: 동전을 100번 던져서 head가 40번 나오고 tail이 60번 나왔다. Head가 나올 확률 μ_1과 tail 이 나올 확률 μ_2 Bayesian learning 으로 추정하려고 한다. μ_i 의 분포는 Dirichlet 분포를 따른다. Prior 분포 $p(\mu_1\,\mu_2|\alpha_1,\alpha_2)$ 에서 hyper-parameter 가 α_1=10, $\alpha_2=20$ 로 주어 졌을 때, μ_1, μ_2 를 추정하면 어떤 값인지 고르시오.

(1) μ_1=2/5, $\mu_2=3/5$ (2) μ_1=5/13, $\mu_2=8/13$ (3) μ_1=6/13, $\mu_2=7/13$

A: (2)

$$\mu_1=\frac{c_1+\alpha_1}{\sum_i(c_i+\alpha_i)}=\frac{40+10}{130}=\frac{5}{13}$$
$$\mu_2=\frac{c_2+\alpha_2}{\sum_i(c_i+\alpha_i)}=\frac{60+20}{130}=\frac{8}{13}$$

Quiz 7: 14, 15 강 퀴즈

Q: Bayesian learning 에서는 특정 클래스의 likelihood distribution 이 parametric pdf (or pmf)를 갖는다고 가정한다.

A: 참

Q: Bayesian learning 에서는 특정 클래스의 likelihood pdf의 parameter 를 미지의 상수로 간주하고 그 클래스로부터 관측된 sample 집합 $D = \{x_i \mid i = 1, \ldots, n\}$ 으로 부터 parameter를 추정한다.

A: 거짓. likelihood pdf의 parameter 를 random variable로 간주하고 parameter이 분포를 추정한다.

Q: Bayesian learning 에서는 likelihood pdf 가 multinomial 분포의 경우, posterior 확률을 Bayes rules로 구해야 하기 때문에 prior 와 likelihood 가 서로 곱해져야 해서 prior 와 likelihood 는 같은 확률 분포를 가져야 한다.

A: 거짓, multinomial의 conjugate prior 분포는 Dirichlet 분포임.

Q: Likelihood 의 분포가 Gaussian 일 때, Bayesian learning 으로 학습된 posteriori 분포는 Gaussian 분포이며, 학습 샘플의 수가 증가할 수록, 이 posteriori 분포의 variance 는 likelihood variance의 참값으로 수렴하게 된다.

A: 거짓. posteriori 분포의 variance는 0으로 수렴하게 된다.

Q: Bayesian learning 에서 특징을 나타내는 랜덤변수와 likelihood의 매개변수는 종속이므로 이 joint 확률을 maginalization 으로 추정 likelihood를 구한다. 이를 위해 Bayesian learning 으로 학습된 매개변수의 분포를 이용하여 total probability 를 구함으로 maginalization 을 수행한다.

A: 참

Q: 특징 x의 분포가 Gaussian 일 때, training data로 부터 Maximum Likelihood Estimation 으로 분포를 구하기 위해서는 training data의 mean과 variance를 구하면 된다.

A: 참

Quiz 7: 16, 17 강 퀴즈

Q:

Non-parametric density estimation 을 위해 histogram 을 이용하여 구할 수 있다. 이를 위해 feature 공간을 bin으로 분할 하였을 때, n 개의 관측 샘플 중 k 개의 샘플이 특정 bin에 속하면 그 bin의 probability density function의 추정 값은 k/n 이 된다.

A: 거짓.

k/n 은 해당 bin에 샘플이 관측될 확률이며 probability density function은 그 확률을 bin의 volume로 나눈 값 $k/(nV)$이 된다.

Q:

Histogram 을 이용한 non-parametric density estimation을 위해서, 차원의 저주 문제를 해결하기 위해, sample 이 수를 점점 늘리면서 추정하는 방법을 쓴다. 이를 위해서 샘플 수가 커질 수록 vin의 크기를 줄여 나가는 방법을 쓴다. 그리고 샘플수가 무한대로 증가할 때, 추정 density function이 참값으로 수렴하기 위해서는, 샘플 수 n이 무한히 증가하면 bin의 크기는 0으로 수렴하고, 해당 bin에서 관측되는 샘플 k_n 은 무한대로 증가해야 하며, k_n/n 은 확률 p로 수렴해야 한다.

A: 거짓.

k_n/n 은 0으로 수렴해야 한다.

Q:

Parzen window 방법은 중심이 테스트 샘플 특징 x인 윈도우 함수 $\phi(x)$ 를 사용하여 pdf를 추정하는 non-parametric 방법입니다. pdf는 샘플 특징이 해당 윈도우 내에 있는 관측 샘플수 k_n 을 다음과 같이 산출한다. $k_n = \sum_{i=1}^{n} \phi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)$, 여기서 h_n 은 샘플 수가 커질수록 커지도록 설계된다.

A: 거짓.

h_n 은 샘플 수가 커질수록 작아지도록 설계된다.

Q:

Parzen window 방식의 pdf 추정법은 모든 관측 샘플을 중심으로 하는 window function을 중첩(superposing)하여 pdf를 추정한다. 하지만 너무 많은 window function을 사용하여 효율적이지 않다. 이를 극복하는 방법으로 클러스터마다 Gaussian kernel을 window function 으로 사용하는 방법이 Gaussian Mixture Model(GMM)이다.

A: 참.

Q:

Parzen window 방식은 모든 영역에서 동일한 윈도우 볼륨을 가지므로 고밀도 영역과 저밀도 영역에 관계없이 추정 해상도가 일정하지 않을 수 있다. 이 문제를 해결하기 위한 k_n-NN 방법은 테스트 특징 x 에 인접한 k_n 개의 관측 샘플을 선택하여 이 들을 포함하는 volume V_n 을 계산하여 pdf 추정치 $p_n(x)$를 구한다. 이는 확률 밀도가 높은 영역은 V_n dl 작아져서 효과적인 pdf추정이 가능하다.

A: 참.

Q:

특징 x 부근에 10 개의 라벨이 있는 샘플을 선택하였다. 이중 4개가 class w_1 에 속하고 6개가 class w_2 에 속한다. 이를 기반으로 특징 x 가 class w_1 속할 확률을 10-nearest neighbor classifier로 구하면 2/5 가 된다.

A: 거짓.

Posteriori prob. $p(w_1|x) = \frac{2}{5}$이므로 Nearest neighbor classifier 로 판단할 때, 2/5이고, 10-nearest neighbor classifier로 특징 x 가 class w_1 속할 확률을 구하면, $p(w_1|\{x_1, \dots, x_{10}\}) = \sum_{i=6}^{10} (^2/_5)^i \, (^3/_5)^{10-i}$이다.

Q:

Gaussian Mixture Model(GMM)로 pdf를 추정하기 위해서는 Gaussian kernel의 매개변수인 각 cluster의 mean 과 variance , 그리고 Gaussian kernel 의 비중(weight) 을 추정해야 하며, Gaussian kernel 의 가중합으로 나타내는 likelihood를 최대로 하는 maximum likelihood estimation 방식으로 구한다. 이를 위해 log likelihood 를 구한 후 Gradient 를 0으로 하여 구한다.

A: 거짓.

각 cluster를 나타내는 latent 변수를 marginalization 하기 위해, 샘플이 각 kernel에 속할 확률을 구해서 log likelihood의 expectation 을 구한 후 Gradient 를 0으로 하여 구한다.

Quiz 7: 18, 19 강 퀴즈

Q:

적분으로 expectation 을 구할 때, 확률 분포에 따라 적분을 구하기 힘든 경우에 Monte Carlo 방법을 사용할 수 있다. 이 방법은 확률 분포로 부터 샘플링하여 적분 대신 합산에 의한 평균값으로 expectation의 근사값을 구하는 것이다.

A: 참.

Q:

특정 확률 분포 $P(x)$로 부터 샘플링하는 것이 어려울 경우 importance sampling 방법을 사용한다. 경우 importance sampling 방법은 샘플링가능한 분포 $Q(x)$ 를 이용하여 샘플링하는 대신, importance weight $\frac{Q(x)}{P(x)}$ 를 대상 값에 곱해서 보정해준다.

A: 거짓.

importance weight은 $\frac{P(x)}{Q(x)}$ 임.

Q:

Importance sampling 을 위한 분포 $Q(x)$ 가 원래 분포 $P(x)$ 와 차이가 많이 날 경우, $Q(x)$ 가 바로 전 상태 x' 에 의해 정해지는 adaptive 분포 $Q(x|x')$ 를 이용하여 샘플링하는 방법이 Markov Chan Monte Carlo 방법이다.

A: 참.

Q:

Restricted Boltzmann machine의 i-th visible neuron 과 j-th hidden neuron 간의 연결가중치 w_{ij}를 학습할 때, 이 두 뉴런간의 firing rate 의 곱의 기대치 $\langle v_i h_j \rangle_{model}$ 를 필요로 한다. 이는 모르는 값이므로 기 학습된 연결가중치에 기반하여 Markov Chain Monte Carlo 방법으로 반복하여 추정하는 방식으로 구한다.

A: 참.

Q:

Markov Chan Monte Carlo 방법으로 샘플링할 때, 샘플의 acceptance probability 는 $A(x'|x) = \min\left(1, \frac{P(x)/Q(x|x')}{P(x')/Q(x'|x)}\right)$로 주어진다. .

A: 거짓.

$A(x'|x) = \min\left(1, \frac{P(x')/Q(x'|x)}{P(x)/Q(x|x')}\right)$ 로 주어짐.

Q:

Gibbs Sampling 은 graphical model에서 서로 연결된 다수의 random variable 들을 반복하여 inference 할 때, 한번 수를 제외하고 나머지 변수들은 전에 inference 된 값으로 관측되었다고 가정하여 condition으로 주고 남은 한 변수만을 inference 하는 방법이다. 이 inference 변수들을 돌아가면서 수렴할 때 까지 반복한다. 이 때 acceptance probability 는 1이 된다.

A: 참.

Q:

Boltzmann 분포 $P(x)$ 는 에너지 $E(x)$ 에 지수함수적으로 역비례하는 확률분포로, 평균에너지가 같을 때, entropy를 최소로 하는 분포이다.

A: 거짓.

entropy를 최대로 하는 분포임.

Q:

Boltzmann machine은 fully connected neural network을 지도 학습하는 모델로서 동시에 발화(firing)하는 neuron 들 간의 가중치 들이 강화되는 모델이며, Error Backpropagation learning rule을 설명할 수 있는 모델이다.

A: 거짓.

비지도학습 모델이며, 신경생물학에서 발견된 Hebbian learning mechanism을 설명할 수 있는 모델임.

Q:

Boltzmann machine에서 특정 뉴런의 발화(firing) 정도를 나타내는 랜덤 변수를 x_i하고 하고 연결된 나머지 뉴런들의 발화(firing) 정도를 나타내는 랜덤 변수들의 집합을 x_{-i} 로 정의하면, $P(x_i = 1|x_{-i})$ 는 다른 뉴런들의 발화에 의해 x_i 가 발화할 확률을 의미한다. 그리고 이 확률은 sigmoid activation을 갖는 뉴런모델 $\sigma(\sum_{j\neq i} w_{ij}x_j + \theta_i)$와 같이 주어진다.

A: 참.

Q:

Restricted Boltzmann machine은 hidden layer와 visible layer를 분리시키고 같은 layer에 속한 뉴런들 간 연결은 없는 모델이다. 이 모델의 학습은 visible neuron 들에서 관측된 firing rate v의 Boltzmann probability $P(v)$를 최대로 하도록 학습한다.

A: 참.