기계학습 Assignment1

2023 2학기

1 (20 points) 본 문항에서는 Bishop's 교재의 Appendix C에 있는 유용한 Matrix Derivatives 공식을 유도하고자 한다. Eq. C.17에 처럼, scalar $x = f(\mathbf{a})$에 대한 벡터 $\mathbf{a}$의 derivative는 다음과 같이 정의된다.

$$\left(\frac{\partial x}{\partial \mathbf{a}}\right)_i = \frac{\partial x}{\partial a_i} \tag{1}$$

Eq. C.18에서 처럼, vector $\mathbf{b} = f(\mathbf{a})$에 대한 벡터 $\mathbf{a}$의 derivative는 다음과 같이 정의된다.

$$\left(\frac{\partial \mathbf{b}}{\partial \mathbf{a}}\right)_{ij} = \frac{\partial b_i}{\partial a_j} \tag{2}$$

다음 matrix derivative 식을 유도하시오.

(i)

$$\frac{\partial}{\partial \mathbf{x}}\left(\mathbf{a}^T\mathbf{x}\right) = \mathbf{a} \tag{3}$$

(ii)

$$\frac{\partial}{\partial \mathbf{x}}\left(\mathbf{x}^T\mathbf{x}\right) = 2\mathbf{x} \tag{4}$$

(iii)

$$\frac{\partial}{\partial \mathbf{x}}\left(\mathbf{x}^T\mathbf{A}\mathbf{x}\right) = 2\mathbf{A}\mathbf{x} \tag{5}$$

(iv) 다음 식 (Eq. C.20)을 유도하고,

$$\frac{\partial}{\partial \mathbf{x}}\left(\mathbf{A}\mathbf{B}\right) = \frac{\partial \mathbf{A}}{\partial \mathbf{x}}\mathbf{B} + \mathbf{A}\frac{\partial \mathbf{B}}{\partial \mathbf{x}} \tag{6}$$

위의 식으로 부터 square matrix $\mathbf{A}$에 대해서 아래 식 (Eq. C.21)을 유도하시오.

$$\frac{\partial}{\partial x}\left(\mathbf{A}^{-1}\right) = -z\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x}\mathbf{A}^{-1} \tag{7}$$

(v) square matrix $\mathbf{A}$에 대해서, 다음 식 (Eq. C.28)을 유도하시오.

$$\frac{\partial}{\partial \mathbf{A}} ln|\mathbf{A}| = \left(\mathbf{A}^{-1}\right)^T \tag{8}$$

2 (15 points) 다음 물음에 답하시오.

(i) 기계학습에서 overfitting이란 무엇인가?

(ii) underfitting이란 무엇인가?

(iii) 기계학습에서 model fitting의 최종 목적은 무엇인가?

(iv) Model complexity와 overfitting, underfitting간의 일반적인 관계를 설명하고, overfitting에 빠지지 않도록 하는 방법 2개정도를 든다면 무엇인가? 각 방법에 대해 간략히 기술하시오.

3 (30 points) 선형 회귀 (linear regression)를 위한 학습 데이터셋 $\mathcal{D} = \{(\mathbf{x}_1, t_1), \cdots, (\mathbf{x}_N, t_N)\}$가 주어지고, 회귀 함수 $y(\mathbf{x}; \theta)$는 $\theta = \{\mathbf{w}, w_0\}$를 파라미터로 하는 다음 함수로 정의된다고 하자.

$$y(\mathbf{x}; \theta) = \mathbf{x}^T\mathbf{w} + w_0$$

다음 물음에 답하시오.

(i) 위의 선형 회귀 함수의 파라미터를 구하기 위해 다음과 같이 square error로 $Loss$를 정의하자.

$$Loss_{\mathcal{D}}(\mathbf{w}, w_0) = \frac{1}{2}\sum_{i=1}^{N}\|y(\mathbf{x}_i; \theta) - t_i\|^2 = \|\mathbf{y} - \mathbf{t}\|^2 \tag{9}$$

여기서, $\mathbf{y}$와 $\mathbf{t}$는 N-dimensional column vectors들로 각각 다음과 같이 정의된다.

$$\mathbf{y} = [y(\mathbf{x}_1; \theta), \cdots, y(\mathbf{x}_N; \theta)]^T \tag{10}$$

$$\mathbf{t} = [t_1, \cdots, t_N]^T \tag{11}$$

다음 parameters에 대한 derivative (gradient)를 구하시오

$$\frac{\partial}{\partial \mathbf{w}}(Loss_{\mathcal{D}}(\mathbf{w}, w_0)) = \tag{12}$$

$$\frac{\partial}{\partial w_0}(Loss_{\mathcal{D}}(\mathbf{w}, w_0)) = \tag{13}$$

(ii) 위의 gradient를 다음 design matrix $\mathbf{X}$를 이용하여 정리하시오.

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1{}^T \\ \vdots \\ \mathbf{x}_N{}^T \end{bmatrix} \tag{14}$$

(iii) 앞서 유도한 gradient식을 이용해서, $\mathbf{w}$와 w_0에 대한 analytic solution을 구하시오.

(iv) 위에서 구한 analytic solution이 **global minimum**임을 증명하시오.

(v) 앞서 유도한 gradient식을 이용해서, $\mathbf{w}$와 w_0에 대한 **gradient descent** 방식을 통한 update식을 기술하시오.

4 (30 points) 앞 문항에서 정의된 bias parameter w_0를 포함하도록 extended parameter vector $\tilde{\mathbf{w}}$를 다음과 같이 정의하자.

$$\tilde{\mathbf{w}} = [w_0; \mathbf{w}^T]^T \tag{15}$$

여기서 ;는 concatenation 연산자로 간주하였다. Bias term의 포함으로 차원이 늘어난 parameter벡터에 맞게, extended feature vector $\tilde{\mathbf{x}}$를 다음과 같이 정의하자.

$$\tilde{\mathbf{x}} = [1; \mathbf{x}^T]^T \tag{16}$$

또한, 회귀 함수 $y(\mathbf{x}; \theta)$를 extended feature vector를 취하는 다음 함수로 정의하자.

$$y(\tilde{\mathbf{x}}; \theta) = \tilde{\mathbf{x}}^T \tilde{\mathbf{w}}$$

다음 물음에 답하시오.

(i) 위의 확장된 회귀 함수의 파라미터로 $Loss_{\mathcal{D}}(\tilde{\mathbf{w}})$ 및 Extended design matrix $\tilde{\mathbf{X}}$를 재정의하고, $\tilde{\mathbf{w}}$에 대한 analytic soultion을 얻으시오.

(ii) Regularization을 위해서 다음 항 $\|\tilde{\mathbf{w}}\|^2$을 추가하여 $Loss$를 다음과 같이 재정의하였다.

$$Loss_{\mathcal{D}}(\tilde{\mathbf{w}}) = \|\mathbf{y} - \mathbf{t}\|^2 + \lambda \|\tilde{\mathbf{w}}\|^2 \tag{17}$$

위의 regularized 버전의 $Loss$에 대한 $\tilde{\mathbf{w}}$의 Gradient를 구하시오.

(iii) Regularized loss를 최소화하기 위한 $\tilde{\mathbf{w}}$의 analytic solution을 구하시오.

(iv) Regularized loss를 최소화하기 위한 $\tilde{\mathbf{w}}$의 gradient descent solutio을 구하시오.

(v) 위와 같은 regularization의 효과는 무엇인가? 기계학습에서 알려진 이슈인 overfitting, generalization과 연관지어 설명하시오.

5 (10 points) X, Y를 랜덤 변수 (random variable)일때, 확률론 (probability theory)의 기본 규칙인 다음 sum rule과 product rule을 이용하여 베이지안 정리 (Bayesian theorm) (Bishop's Eq. 1.12)을 유도하시오.

$$\begin{array}{rl} \textbf{sum rule} & p(X) = \sum_Y P(X, Y) \\ \textbf{product rule} & p(X, Y) = p(Y|X)p(X) \end{array}$$

베이지안 정리는 다음과 같다.

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \tag{18}$$

6 (20 points) Decision 문제는 주어진 feature vector $\mathbf{x}$에 대한 범주를 결정하는 문제로, $p(\mathcal{C}_k|\mathbf{x})$를 구하거나 discriminant function $f(\mathbf{x})$를 구하는 접근법으로 나뉠 수 있다. 다음 물음에 답하시오.

(i) $p(\mathcal{C}_k|\mathbf{x})$을 구하는 방법은 다시 **generative model**을 이용한 방법, **discriminant model**에 기반한 방식으로 나뉘는데, 이들 두 방법이 가각 무엇인지 설명하시오.

(ii) 앞서의 **generative model**과 **discriminant model**에 기반한 방법의 장단점이 무엇인지 상세히 서술하시오.

(iii) $f(\mathbf{x})$은 $\mathbf{x}$를 class로 매핑하는 **discriminant function**으로 $p(\mathcal{C}_k|\mathbf{x})$를 통하지 않고 직접 구하는 방식의 예를 기술하시오.

(iv) **discriminant function**에 기반한 방법대비 $p(\mathcal{C}_k|\mathbf{x})$를 통하여 decision을 수행하는 probabilistic 방법의 장점을 기술하시오.

7 (15 points) 1-dimensional random variable에 대해 평균이 μ이고 분산이 σ^2일때, **differential entropy**를 최대로 하는 probability distribution p가 Gaussian distribution임을 증명하시오 (Bishop교재 Eq. 1.109를 유도하면 된다)

8 (30 points) parameter μ를 갖는 Bernoulli distribution은 다음과 같이 정의 된다 (Bishop의 Eq. 2.2)

$$Bern(x|\mu) = \mu^x(1-\mu)^{(1-x)} \tag{19}$$

μ가 unknown이고, N개의 observed 데이터 $\mathcal{D} = \{x_1, \cdots, x_N\}$가 주어질때의 likelihood는 다음과 같다.

$$p(\mathcal{D}|\mu) = \prod_{i=1}^{N} Bern(x_i|\mu) \tag{20}$$

다음 물음에 답하시오.

(i) 위의 likelihood Eq. 20를 최대로 하는 μ에 대한 식을 유도하시오.

(ii) μ를 랜덤 변수화 하고 이에 대한 prior distribution를 위해 mu가 다음 Beta distribution을 따른다고 가정하자.

$$p(\mu) = p(\mu|a,b) = Beta(\mu|a,b) \tag{21}$$

parameter a, b가 주어질때 beta distribution의 평균과 분산에 대한 식을 유도하시오.

(iii) Conjugacy property란 무엇인가? μ에 대한 prior distribution을 위해 beta distribution를 사용한 이유는 무엇인가?

(iv) 위와 같이 N개의 데이터 $\mathcal{D} = \{x_1, \cdots, x_N\}$가 주어질때, μ에 대한 posterior distribution는 다음과 같다.

$$p(\mu|\mathcal{D}) \propto p(\mathcal{D}|\mu)p(\mu|a,b) \tag{22}$$

위의 posterior가 beta distribution을 따름을 보이고,해당 beta분포의 parameter에 대한 정확한 식을 유도하시오.

(v) 위의 posterior 분포를 이용해서 다음과 같이 $x = 1$일 확률인 **predictive distribution**를 구하고자 한다.

$$p(x = 1|\mathcal{D}) = \int_0^1 p(x = 1|\mu)p(\mu|\mathcal{D})d\mu \tag{23}$$

위의 식을 hyper parameter a, b 및 dataset의 N개의 관측값인 x_i의 함수로 간결하기 기술하고, 이를 간단히 유도하시오.

9 (30 points) parameter vector $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_K)^T$를 갖는 Multinoulli distribution은 다음과 같이 정의 된다 (Bishop의 Eq. 2.26)

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k} \tag{24}$$

N개의 one-hot vector로 구성된 관측 데이터 $\mathcal{D} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$가 주어질때, likelihood는 다음과 같다 (Bishop의 Eq. 2.29).

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \mu_k^{x_{nk}} \tag{25}$$

다음 물음에 답하시오.

(i) 위의 likelihood Eq. 25를 최대로 하는 μ_k에 대한 식을 유도하시오.

(ii) $\boldsymbol{\mu}$에 대한 conjugate prior분포로 다음 Dirichlet distribution을 가정하자.

$$p(\boldsymbol{\mu}) = Dir(\boldsymbol{\mu}|\boldsymbol{\alpha}) \tag{26}$$

이때 $\boldsymbol{\alpha}$는 $(a_1, \cdots, a_K)^T$로 구성된 hyperparameter vector를 가리킨다. conjugacy에 의하여, 위의 $\mathcal{D}$에 대한 $\boldsymbol{\mu}$의 posterior distribution은 Dirichlet 분포를 따르며, 이때의 updated hyper parameter를 구하고 이를 유도하시오.

(iii) 위의 posterior 분포를 이용해서 다음과 같이 $x_k = 1$일 확률인 posterior **predictive distribution**를 구하고자 한다.

$$p(x_k = 1|\mathcal{D}) = \int p(x_k = 1|\boldsymbol{\mu})p(\boldsymbol{\mu}|\mathcal{D})d\boldsymbol{\mu} \tag{27}$$

위의 predictive distribution을 prior의 hyperparameter $\{a_k\}_{k=1}^K$과 N개의 관측벡터 $\{\mathbf{x}_i\}_{i=1}^N$의 식으로 간결히 하고, 이를 유도하시오.

10 (10 points) D-dimensional vector $\mathbf{x}$에 대한, multivariate Gaussian distribution은 아래와 같이 정의된다.

$$\mathcal{N}(\mathbf{x}|\mu, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\} \tag{28}$$

다음 물음에 답하시오.

(i) Covariance matrix $\boldsymbol{\Sigma}$에 대한 eigendecomposition결과가 다음과 같다고 하자.

$$\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T \tag{29}$$

여기서 $\boldsymbol{\Lambda}$는 $\boldsymbol{\Sigma}$의 eignvalue λ_i로 구성된 대각행렬 (diagonal matrix)이다.

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \cdots & \cdots & \vdots \\ 0 & 0 & \cdots & \lambda_D \end{bmatrix} \tag{30}$$

이때 아래 교재의 Eq. 2.52의 변환을 이용하여,

$$\mathbf{y} = \mathbf{U}(\mathbf{x} - \boldsymbol{\mu}) \tag{31}$$

위의 multivariate Gaussian distribution (Eq. 28)를 Bishop교재의 Eq. 2.56와 같이 D개의 독립 (independent)된 univariate Gaussian distributions의 곱으로 표현되는 과정을 유도하시오.

11 (10 points) Bishop교재의 Partitioned matrix에 대한 다음 inverse식 (Eq. 2.76) 이 정당함을 유도하시오.

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M} & -\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{M} & -\mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \end{pmatrix} \tag{32}$$

이때, $\mathbf{M}$은 위의 좌측 행렬에 대한 Schur complement로 다음과 같이 정의된다.

$$\mathbf{M} = \left(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1} \tag{33}$$

12 (20 points) D-dimensional vector $\mathbf{x}$에 대한 joint Gaussian distribution $\mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma})$이 주어지고, precision matrix 는 covariance matrix의 inverse로 $\mathbf{\Lambda} = \mathbf{\Sigma}^{-1}$형태로 주어졌다고 하자. 입력 차원을 두 block으로 나누어,

$$\begin{aligned} \mathbf{x} &= \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \quad (34) \\ \mathbf{\Sigma} &= \begin{pmatrix} \mathbf{\Sigma}_{aa} & \mathbf{\Sigma}_{ab} \\ \mathbf{\Sigma}_{ba} & \mathbf{\Sigma}_{bb} \end{pmatrix}, \\ \mathbf{\Lambda} &= \begin{pmatrix} \mathbf{\Lambda}_{aa} & \mathbf{\Lambda}_{ab} \\ \mathbf{\Lambda}_{ba} & \mathbf{\Lambda}_{bb} \end{pmatrix} \quad (35) \end{aligned}$$

로 할때, 다음 conditional distribuation에 대한 식 (Eq. 2.96)

$$\begin{aligned} p\left(\mathbf{x}_a|\mathbf{x}_b\right) &= \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_{a|b}, {\mathbf{\Lambda}_{aa}}^{-1}\right) \\ \boldsymbol{\mu}_{a|b} &= \quad (36) \end{aligned}$$

Marginal distribution에 대한 식 (Eq. 2.98)

$$p\left(\mathbf{x}_a\right) = \mathcal{N}\left(\mathbf{x}_a|\boldsymbol{\mu}_a, \mathbf{\Sigma}_{aa}\right)$$

에서 $\boldsymbol{\mu}_{a|b}$을 완성하고, 위의 두 conditional distribution 및 marginal distribution 식을 유도하시오.

13 (20 points) $\mathbf{x}$에 대한 marginal distribution과, $\mathbf{x}$가 주어질때의 $\mathbf{y}$에 대한 conditional distribution이 다음과 같이 Gaussian을 따른다고 하자.

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}\left(\boldsymbol{\mu}, \mathbf{\Lambda}^{-1}\right) \\ p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}\left(\mathbf{A}\mathbf{x} + \mathbf{b}, \boldsymbol{L}^{-1}\right) \end{aligned}$$

이때 $\mathbf{y}$에 대한 marginal distribution과 $\mathbf{y}$가 주어질때의 $\mathbf{x}$에 대한 conditional distribution의 다음 식을 완성하시오 (Bishop Eq. 2.115 와 2.116).

$$\begin{aligned} p(\mathbf{y}) &= \\ p(\mathbf{x}|\mathbf{y}) &= \end{aligned}$$

또한, Bishop 교재를 참조하면서 위의 $p(\mathbf{y})$와 $p(\mathbf{x}|\mathbf{y})$의 공식을 유도하시오.

14 (15 points) Multivariate Gaussian distribution을 따르는 관측 데이터가 Design matrix로 $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)^T$로 주어졌다고 하자. 이에 대한 log-likelihood를 최대화 하는 평균과 분산의 MLE (maximum likelihood estimation)이 다음과 같음을 증명하시오 (교재 Eq. 2.121- 2.122).

$$\begin{aligned} \boldsymbol{\mu}_{ML} &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \\ \mathbf{\Sigma}_{ML} &= \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{x}_n - \boldsymbol{\mu}_{ML}\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_{ML}\right)^T \end{aligned}$$

15 (15 points) 분산이 σ^2이 정해졌으나 μ가 unknown인 Gaussian distribution에 대한 Bayesian inference를 위해, 총 N개의 1-dimensional observed samples들이 Design vector인 $\mathbf{X} = (x_1, \cdots, x_N)^T$로 주어졌다고 하자. μ에 대한 Bayesian inference를 위해 prior distribution으로 다음 Gaussian을 따른다고 하자.

$$p(\mu) = \mathcal{N}(\mu_0, {\sigma_0}^2) \tag{37}$$

Bishop교재의 Eq. 2.116를 이용하여, μ에 대한 다음 posterior distribution의 파라미터 μ_N와 ${\sigma_N}^2$을 유도하시오 (Bishop교재 Eq. 2.141-142).

$$p(\mu|\mathbf{X}) = \mathcal{N}(\mu|\mu_N, {\sigma_N}^2) \tag{38}$$

16 (15 points) 선형 회귀에 대한 Least squares 방식의 해가 Gaussian noise 모델의 maximum likelihood estimation를 보이고자 한다. 이를 위해 주어진 입력 벡터 $\mathbf{x}$ (bias entry가 포함된 확장된 벡터로 간주)에 대한 regression 값 t에 대한 probability distribution을 다음과 같이 Gaussian을 따른다고 하자.

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}\left(t|\mathbf{w}^T\mathbf{x}, \beta^{-1}\right) \tag{39}$$

N개의 관측 데이터가 Design matrix $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)^T$로, 이에 대한 target regression values들이 $\mathbf{t} = (t_1, \cdots, t_N)$로 주어졌다고 하자. 데이터에 대한 likelihood는 다음과 같다.

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^T\mathbf{x}_n, \beta^{-1}) \tag{40}$$

다음 물음에 답하시오.

(i) 위의 Log-likelihood를 전개하면 Sum-of-squares error항이 포함됨을 유도하시오 (교재 Eq 3.11).

(ii) 위의 Log-likelihood를 최대로 하는 β^{-1}의 식을 유도하고, 해당 값의 의미에 대해서 서술하시오.

17 (15 points) 앞 문항을 확장하여 Bayesian linear regression을 유도하고자 한다. 이를 위해, parameter vector $\mathbf{w}$의 prior distribution이 다음 Gaussian분포를 따른다고 가정하자 (교재 Eq. 3.48).

$$p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0\right) \tag{41}$$

다시 Bishop교재의 Eq. 2.116를 적용하여, $\mathbf{w}$에 대한 다음 posterior distribution의 파라미터 $\mathbf{m}_N$와 $\boldsymbol{S}_N$을 유도하시오 (Bishop교재 Eq. 3.50-51).

$$p(\mathbf{w}|\mathbf{X}, \mathbf{t}) = \mathcal{N}\left(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N\right) \tag{42}$$

18 (30 points) Binary classification에서의 Logistic regression 학습과정을 유도하고자 한다. 입력 벡터 $\mathbf{x}$ 에 대해, class $\mathcal{C}_1$으로 분류할 확률은 linear regression된 값에 logistic sigmoid function을 적용하여 다음과 같이 구한다고 가정한다 (교재 Eq. 4.87)

$$p(\mathcal{C}_1|\mathbf{x}) = \sigma\left(\mathbf{w}^T\mathbf{x}\right) \tag{43}$$

N개의 입력 데이터가 Design matrix $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)^T$로, 이에 대한 target **class** values 이 $\mathbf{t} = (t_1, \cdots, t_N)^T$로 주어진다고 하자. Regression setting과 달리 여기서, $t_n \in \{0, 1\}$로 이진값으로 제한된다. Data likelihood는 다음과 같다.

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}) = \prod_n y_n^{t_n}(1 - y_n)^{1-t_n} \tag{44}$$

여기서, $y_n = p(\mathcal{C}_1|\mathbf{x}_n) = \sigma\left(\mathbf{w}^T\mathbf{x}_n\right)$이다.
다음 물음에 답하시오.

(i) $Loss(\mathbf{w})$를 다음과 같이 negative log-likelihood로 두자.

$$Loss(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}) \tag{45}$$

위의 Loss에 대한 $\mathbf{w}$의 다음 gradient를 구하시오.

$$\nabla_{\mathbf{w}} Loss(\mathbf{w}) = \frac{\partial}{\partial \mathbf{w}} Loss(\mathbf{w}) = \tag{46}$$

Gradient 를 Design matrix $\mathbf{X}$를 이용하여 간결하게 정리하시오.

(ii) 위의 Loss에 대한 $\mathbf{w}$ Hessian matrix를 구하시오. $f(\mathbf{x})$에 대한 vector $\mathbf{x}$ Hessian matrix $\mathbf{H}$는 다음과 같이 정의된다.

$$\mathbf{H}_{ij} = \frac{\partial f(\mathbf{x})}{\partial x_i x_j}$$

Hessian matrix를 Design matrix $\mathbf{X}$를 이용하여 간결하게 정리하시오.

(iii) $\frac{\partial}{\partial \mathbf{w}} Loss(\mathbf{w}) = \mathbf{0}$가 되는 analytic solution을 구할 수 있는가? 근거를 설명하시오.

(iv) Gradient descent방식에 기반한 $\mathbf{w}$의 update수식을 기술하시오.

19 (20 points) Multiclass Logistic regression 학습방법을 유도하고자 한다. K개의 class중에서 입력 벡터 $\mathbf{x}$ 에 대해, class $\mathcal{C}_k$으로 분류할 확률은 linear regression된 값에 softmax를 적용하여 다음과 같이 얻는다 (교재 Eq. 4.104).

$$p(\mathcal{C}_k|\mathbf{x}) = y_k(\mathbf{x}) = \text{softmax}_k\left(a_{i=1}^K\right) \tag{47}$$

여기서 $a_k = {\mathbf{w}_k}^T\mathbf{x}$이고 softmax_k는 다음과 같이 정의된다.

$$\text{softmax}_k\left(a_{i=1}^K\right) = \frac{exp(a_k)}{\sum_j exp(a_j)} \tag{48}$$

N개의 입력 데이터가 Design matrix $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)^T$로, 이에 대한 **target matrix**는 N개의 one-hot vector로 $\mathbf{T} = (\mathbf{t}_1, \cdots, \mathbf{t}_N)^T$로 주어진다고 하자. 여기서, $\mathbf{t}_i$는 K-dimensional one-hot vector이다. 이때 Data likelihood는 다음과 같다 (교재 Eq. 4.107).

$$p(\mathbf{T}|\mathbf{X}, \mathbf{w}_1 \cdots, \mathbf{w}_K) = \prod_{n=1}^N \prod_{k=1}^K y_{nk}^{t_{nk}} \tag{49}$$

여기서 $y_{nk} = y_k(\mathbf{x})$이다.
다음 물음에 답하시오.

(i) $Loss(\mathbf{w}_1, \cdots, \mathbf{w}_K)$를 다음과 같이 cross entropy (negative log-likelihood)로 두자.

$$Loss(\mathbf{w}_1, \cdots, \mathbf{w}_K) = -\ln p(\mathbf{T}|\mathbf{X}, \mathbf{w}_1 \cdots, \mathbf{w}_K) \tag{50}$$

위의 Loss에 대한 $\mathbf{w}_j$의 다음 gradient를 구하시오.

$$\nabla_{\mathbf{w}_j} Loss(\mathbf{w}_1, \cdots, \mathbf{w}_K) = \frac{\partial}{\partial \mathbf{w}_j} Loss(\mathbf{w}_1, \cdots, \mathbf{w}_K) = \tag{51}$$

Gradient 를 Design matrix $\mathbf{X}$와 target matrix $\mathbf{T}$를 이용하여 간결하게 표현하시오.

제출 방법

각 문항별 수식 유도 및 풀이 과정을 정리한 **문제 풀이 답안지** 파일을 email로 제출해야 한다. (수기 작성시에는 별도 용지의 스캔본 제출 보다는 전자펜으로 답안을 작성하여 생성된 파일로 제출하는 것을 권장함).