기계학습 (2022 년 2 학기)
Assignment #1 – 문제풀이

학번: 이름:

1. **[Nearest Neighbors and the Curse of Dimensionality]** 수업시간에 배웠던 "고차원 공간에서는 대부분의 데이터 샘플들은 서로 멀리 떨어져 있고, 모든 점들 사이의 거리는 대략적으로 같다" 라는 것을 수식적으로 증명해 본다.

   (a) 서로 독립인 두 확률 변수(random variable) X와 Y가 있고 이들은 [0,1] 구간에서 uniform distribution에 따라 sampling되는 값을 가진다. X와 Y 사이의 거리의 제곱으로 정의되는 확률 변수 $Z=(X-Y)^2$의 기대값 $\mathbb{E}[Z]$과 분산 $VAR[Z]$의 값을 구하시오.

   (b) d 차원의 unit cube 공간에서 두 개의 점들을 샘플링한다고 가정하자. d 차원 좌표의 각 성분들은 독립적으로 [0,1] 사이의 값을 uniform distribution에 따라 가진다. 즉 두개의 점을 구성하는 성분들을 각각 d 개의 확률 변수들 $X_1,..,X_d$, $Y_1,..,Y_d$로 나타낼 수 있다. 두 점 사이의 Euclidean 거리 R은 $R=Z_1+\cdots+Z_d$로 나타낼 수 있고 $Z_i=(X_i-Y_i)^2$이다. 기대값과 분산의 성질을 이용하여 $\mathbb{E}[R]=d\cdot\mathbb{E}[z]$, $VAR[R]=d\cdot VAR[Z]$임을 증명하시오.

(c) (b)의 결과가 "고차원 공간에서 대부분의 데이터 샘플들은 멀리 떨어져 있고, 모든 점들 사이의 거리는 대략적으로 같다"를 어떻게 입증하는지 설명하시오.

힌트: 대부분의 데이터 샘플들이 멀리 떨어져 있다는 것은 차원이 커질 수록 $\mathbb{E}[R]$이 커진다는 것을 보인다. 모든 점들 사이의 거리가 대략적으로 같다는 것은 차원이 커질수록 임의의 두 점 사이의 거리 R과 $\mathbb{E}[R]$이 비슷해진다는 것을 보인다. 즉 $\lim_{d\to\infty} VAR[\frac{R}{\mathbb{E}[R]}] = 0$ 을 증명한다.

2. **[Information Theory]** 확률 밀도 함수 p 를 가지는 이산 확률 변수 X 의 엔트로피는 $H(X) = \sum_x p(x)\log_2\left(\frac{1}{p(x)}\right)$로 정의된다. 여기서 $x \in X$이고, X는 일반적으로 유한한 크기의 집합으로 가정한다. $p(x) = 0$이면 $(x)\log_2\left(\frac{1}{p(x)}\right) = 0$으로 가정할 수 있다.

(a) $H(X) \geq 0$ 임을 증명하시오.

Information theory 의 중요한 개념 중의 하나는 두 확률 분포 p와 q의 relative entropy 로서 KL-divergence 라는 이름으로 잘 알려져 있다. 아래와 같이 정의된다.

$$KL(p||q) = \sum_x p(x) \log_2 \frac{p(x)}{q(x)}$$

KL-divergence 는 두 확률 분포 사이의 차이를 측정하기 위해 가장 많이 사용되는 방법 중의 하나로서 정보이론, 기계학습, 통계학 분야에서 널리 활용된다. 모든 x에 대해 $p(x) \geq 0$, $q(x) \geq 0$ 이라고 가정한다. 두 확률 분포가 비슷해지면 KL-divergence 는 작아지고, 같아지면 0 이 된다. KL-divergence 는 symmetric 하지 않고, 삼각 부등식을 만족하지 않기 때문에 엄밀하게 말하면 distance metric 이라고 할 수 없다. 그러나 두 확률 분포의 차이를 측정하기 위해 많이 사용된다.

(b) $KL(p||q) \geq 0$임을 증명하시오.
힌트: Jensen's Inequality 를 활용한다. 즉 함수 $\phi(x)$가 convex 함수일 때는 $\phi(E[x]) \leq E[\phi(x)]$이고 concave 함수인 경우에는 반대 부등호가 성립함.

(c) 두 확률 변수 X와 Y에 대해 Information gain 또는 mutual information $I(Y;X) = H(Y) - H(Y|X)$로 정의된다. 아래의 식을 증명하시오.

$$I(Y;X) = KL\big(p(x,y)||p(x), p(y)\big)$$

$p(x) = \sum_y p(x,y)$는 X의 marginal distribution 을 의미함.

3. **[AdaBoost]** AdaBoost 에서는 weak learner 들이 어려운 학습데이터들에 집중하도록 하기 위해 학습데이터 샘플들에 대한 가중치를 변경한다. 이번 문제에서는 AdaBoost 를 이용해 weak learner 들이 target label 이 $\{-1,+1\}$인 이진분류를 학습하는 것을 고려한다. 수업시간에 배운 것처럼 AdaBoost 알고리즘의 t 번째 iteration 에서 학습되는 weak learner h_t는 아래와 같이 나타낸다.

$$h_t \leftarrow \underset{h\in H}{\operatorname{argmin}} \sum_{i=1}^{N} w_i \mathbb{I}\{h(X^{(i)}) \neq t^{(i)}\}$$

h_t의 분류 오차 err_t는

$$err_t = \frac{\sum_{i=1}^{N} w_i \mathbb{I}\{h(X^{(i)}) \neq t^{(i)}\}}{\sum_{i=1}^{N} w_i}$$

이고 분류기 계수(또는 신뢰도)는 $\alpha_t = \frac{1}{2}\ln\left(\frac{1-err_t}{err_t}\right)$로 나타낸다.

AdaBoost 알고리즘에서는 weak learner h_t가 각 학습데이터 샘플들을 정확하게 분류하는지에 따라 샘플들의 가중치를 변경한다. 데이터 샘플 i의 t+1 번째 iteration 에서 사용될 가중치 w_i'는 아래와 같이 계산된다.

$$w_i' \leftarrow w_i \exp\left(-\alpha_t t^{(i)} h_t(X^{(i)})\right)$$

(a) t+1 번째 iteration 에서의 학습데이터의 가중치 $(w_1', \ldots, w_N')$를 이용하여 weak learner h_t에 적용한다고 가정해 보자. 이 때 h_t의 분류 성능은 $\frac{1}{2}$ 임을 보이시오. 이는 아래의 수식을 증명하는 것과 같다.

$$err_t' = \frac{\sum_{i=1}^{N} w_i' \mathbb{I}\{h(X^{(i)}) \neq t^{(i)}\}}{\sum_{i=1}^{N} w_i'} = \frac{1}{2}$$

(b) 위의 결과는 AdaBoost 알고리즘과 어떻게 연관되는지 간단하게 설명하시오.

4. **[Linear Regression]** 제곱 오차 (squared error) 손실 함수를 이용한 linear regression 은 outlier 에 크게 반응하는 문제점을 가지고 있다. 이를 완화하기 위해 Hubber loss 라는 손실 함수 L_δ 는 hyperparamter δ를 이용하여 다음과 같이 정의한다.

$$L_\delta(y,t) = H_\delta(y-t)$$

$$H_\delta(a) = \begin{cases} \frac{1}{2}a^2, & |a| \le \delta \\ \delta\left(|a| - \frac{1}{2}\delta\right), & |a| > \delta \end{cases}$$

(a) Hubber loss $L_\delta(y,t)$와 제곱 오차 손실함수 $L_{SE}(y,t) = \frac{1}{2}(y-t)^2$를 $t=0$에 대해 그래프로 그리고 Hubber loss 가 outlier 에 더 강인한 이유를 설명하시오.

(b) 다음과 같은 선형 모델을 가정하자.

$$y = \boldsymbol{w}^T\boldsymbol{x} + b$$

이를 이용하여 편미분 $\partial L_\delta/\partial \boldsymbol{w}$과 $\partial L_\delta/\partial \boldsymbol{b}$를 구하시오. 미분값 $H'_\delta(a)$를 먼저 계산하고 이를 이용해 $H'_\delta(y-t)$를 계산하면 좀더 수월하다.

5. **[Linear Regression]** 학습데이터 $\{(x^{(1)}, y^{(1)}), \dots, (x^{(N)}, y^{(N)})\}$와 양의 값을 갖는 가중치 $a^{(1)}, \dots, a^{(N)}$에 대해 weighted least square 문제는 아래와 같이 나타낼 수 있다.

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \boldsymbol{L}(\boldsymbol{w}) = \arg\min_{\boldsymbol{w}} \frac{1}{2}\sum_{i=1}^{N} a^{(i)}\left(y^{(i)} - \boldsymbol{w}^T\boldsymbol{x}^{(i)}\right)^2 + \frac{\lambda}{2}\left||\boldsymbol{w}|\right|^2$$

위의 식에 대한 해는 아래와 같이 나타낼 수 있음을 증명하시오.

$$\boldsymbol{w}^* = (\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}X^T\boldsymbol{A}\boldsymbol{Y}$$

$\boldsymbol{X}$는 수업시간에 정의했던 design matrix, $\boldsymbol{A}$는 성분 $\boldsymbol{A}_{ii} = a^{(i)}$ 인 대각 행렬을 의미한다.
힌트: $\boldsymbol{L}(\boldsymbol{w}) = \frac{1}{2}(Y - XW)^T A(Y - XW) + \frac{\lambda}{2}W^T W$의 행렬식으로 나타낼 수 있다. 이를 전개하여 $\nabla\boldsymbol{L}(\boldsymbol{w}^*) = 0$을 만족하는 $\boldsymbol{w}^*$를 구한다.

6. **[Linear Regression]** 아래의 문제들에 대해 각각 유도 과정을 제시하시오.

(a) Linear regression 에서 least squares 방법을 이용한 loss function $l(\boldsymbol{w})$이 아래와 같이 주어져 있다. 파라메터들에 대한 설명은 강의 자료를 참고한다.

$$l(\boldsymbol{w}) = \sum_{n=1}^{N} \left[t^{(n)} - \left(w_0 + w_1 x^{(n)}\right)\right]^2$$

Least squares 를 이용한 linear regression 문제는 analytical 한 방법으로 최적의 해(즉 $\boldsymbol{w}$)를 구할 수 있다. 벡터 $\boldsymbol{t}$ 와 $\boldsymbol{X}$를 다음과 같이 정의하자.

$$\boldsymbol{t} = \left[t^{(1)}, t^{(2)}, \dots, t^{(N)}\right]^T$$

$$\boldsymbol{X} = \begin{bmatrix} 1, x^{(1)} \\ 1, x^{(2)} \\ \cdots \\ 1, x^{(N)} \end{bmatrix}$$

이로부터 $\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{t}$ 임을 증명하시오. 아래의 자료에서 4.4 절을 읽어보면 풀이에 많은 도움이 된다. (주의: 자료의 내용을 그대로 번역해서 제출하는 것은 인정하지 않음. 반드시 더 자세한 풀이과정이 있어야 함)

http://cs229.stanford.edu/section/cs229-linalg.pdf

(b) Regularization 을 고려하면 위에서 정의된 $l(\boldsymbol{w})$로부터 새로운 loss function $\tilde{l}(\boldsymbol{w})$을 아래와 같이 정의할 수 있다.

$$\tilde{l}(\boldsymbol{w}) = \sum_{n=1}^{N} \left[t^{(n)} - \left(w_0 + w_1 x^{(n)}\right)\right]^2 + \alpha \boldsymbol{w}^T \boldsymbol{w}$$

위의 문제처럼 analytical 한 방법을 적용하면 $\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X} + \alpha \boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{t}$ 와 같이 표현된다. 이를 유도하시오.

7. **[Logistic Regression]** Logistic regression 은 linear regression 과 비교해 overfitting 에 강한 특성을 가지고 있다. 왜 그런지 설명하시오.

8. **[Logistic Regression]** Multi-class logistic regression 에서 주어진 $\boldsymbol{w}$에 대해 training set $\boldsymbol{X}$의 likelihood $p(\boldsymbol{T}|\boldsymbol{w_1},\dots,\boldsymbol{w_k})$는 아래와 같이 정의된다. 수식에 포함된 파라메터들에 대한 자세한 설명은 강의 자료를 참고한다.

$$p(\boldsymbol{T}|\boldsymbol{w_1},\dots,\boldsymbol{w_k}) = \prod_{n=1}^{N}\prod_{k=1}^{K} p\left(C_k|\boldsymbol{x}^{(n)}\right)^{t_k^{(n)}} = \prod_{n=1}^{N}\prod_{k=1}^{K} y_k^{(n)}\left(\boldsymbol{x}^{(n)}\right)^{t_k^{(n)}}$$

여기서 $p(C_k|\boldsymbol{x})$는 softmax function 을 이용한다. 즉 $p(C_k|\boldsymbol{x}) = y_k(\boldsymbol{x}) = \frac{\exp(z_k)}{\sum_j \exp(z_j)}$ 로 나타낼 수 있다. Negative log-likelihood 를 loss function 으로 사용하면 이는 아래의 cross-entropy 함수가 된다.

$$E(\boldsymbol{w_1},\dots,\boldsymbol{w_k}) = -\log\left(p(\boldsymbol{T}|\boldsymbol{w_1},\dots,\boldsymbol{w_k})\right) = -\sum_{\mathrm{n}=1}^{\mathrm{N}}\sum_{\mathrm{k}=1}^{\mathrm{K}} t_k^{(n)} \log\left[y_k^{(n)}\left(\boldsymbol{x}^{(n)}\right)\right]$$

$E(\boldsymbol{w_1},\dots,\boldsymbol{w_k})$를 최소화하는 $\boldsymbol{w}$를 구하기 위해서는 gradient descent 를 이용한다. 이를 위해서는 $\boldsymbol{w}$의 각 성분들에 대한 $E(\boldsymbol{w_1},\dots,\boldsymbol{w_k})$의 편미분값, 즉 $\frac{\partial E}{\partial w_{k,i}}$를 구해야 하는데 이는 아래와 같이 chain rule 을 이용하여 구할 수 있다.

$$\frac{\partial E}{\partial w_{k,i}} = \sum_{\mathrm{n}=1}^{\mathrm{N}}\sum_{\mathrm{k}=1}^{\mathrm{K}} \frac{\partial E}{\partial y_j^{(n)}} \cdot \frac{\partial y_j^{(n)}}{\partial z_k^{(n)}} \cdot \frac{\partial z_k^{(n)}}{\partial w_{k,i}} = \sum_{\mathrm{n}=1}^{\mathrm{N}} \left(y_k^{(n)} - t_k^{(n)}\right) \cdot x_i^{(n)}$$

$\frac{\partial E}{\partial w_{k,i}}$가 위의 식과 같다는 것을 증명하시오. 수업시간에 배운 것과 같이 $\frac{\partial E}{\partial y_j^{(n)}}, \frac{\partial y_j^{(n)}}{\partial z_k^{(n)}}, \frac{\partial z_k^{(n)}}{\partial w_{k,i}}$ 를 각각 구한 후 이를 위의 식에 대입하여 오른쪽 항이 나오도록 정리한다.