기계학습 기말고사 (2019년도 2학기)

2019년 12월 16일 11:00-13:00

학번: 이름:

---

1. 아래의 문장들 중 맞는 것들을 고르시오.
   A. 은닉 레이어들이 없고 출력 레이어에서 logistic activation function를 사용하지 않는 신경망은 선형 결정 평면(linear decision boundary)을 출력하게 된다. ( )
   B. Gaussian Naïve Bayes 분류기는 임의의 클래스에 대해 입력 특징들의 공분산 행렬은 대각 행렬을 가정한다. ( )
   C. Gaussian Discriminant Analysis는 완전한 공분산 행렬(full covariance matrix)를 사용할 때 이차식 형태의 결정 평면(quadratic decision bound)를 가진다. ( )

2. K-mean 알고리즘의 목적함수를 한 문장으로 설명하시오.

3. μ와 β를 모수로 갖는 Laplace 분포는 아래와 같이 정의된다.

$$Laplace(w; \mu, \beta) = \frac{1}{2\beta} exp\left(-\frac{|w-\mu|}{\beta}\right)$$

   Bayesian linear regress의 변형된 형태를 고려한다. 즉 가중치 벡터 w는 평균이 0인 라플라스 분포를 따르는 독립적인 차원들로 구성되어 있으며 아래와 같이 모수 β를 공유한다.

$$w_j \sim Laplace(0, \beta) \qquad t \mid \boldsymbol{w} \sim \mathcal{N}(t\,;\, \mathbf{w}^{\mathrm{T}}\,\psi(\mathbf{x}), \sigma)$$

   Gaussian 분포는 $\mathcal{N}(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$로 정의된다.

   A. 학습 데이터 집합 $\{(x^{(i)}, t^{(i)})\}_{i=1}^{N}$를 가지고 있다고 가정하자. MAP(Maximum A Posteriori)를 이용해 $\boldsymbol{w} = \{\boldsymbol{w_j}\}$를 구하려고 할 때 필요한 목적함수를 $\boldsymbol{w_j}$와 $x^{(i)}, t^{(i)}$의 식으로 나타내시오.

(힌트: $\hat{\theta}_{MAP} = \underset{\theta}{argmax}\, P(\theta|D) = \underset{\theta}{argmax}\, P(\theta, D) = \underset{\theta}{argmax} P(\theta)P(D|\theta) = \underset{\theta}{argmax}\, log\, P(\theta) + log\, P(D|\theta)$ 이용)

   B. 위에서 기술한 목적 함수에 대해 Laplace prior를 이용한 MAP의 해는 Gaussian prior를 이용한 MAP의 해와 어떻게 다른지 한 문장으로 설명하시오.

4. 두 개의 단변수 Gaussian 분포의 조합으로 구성된 Gaussian Mixture model (GMM)을 EM 알고리즘으로 최적화하려고 한다. 수업시간에 배웠던 GMM과 다르게 두 Gaussian 분포의 평균은 μ으로 동일하지만, 서로 다른 표준 편차 σ_k를 가지고 있다. 두 분포 중의 하나를 가리키는 확률 변수는 $z \in \{0,1\}$로 정의한다. GMM은 아래와 같이 나타낸다.

$$z \sim Bernoulli(\theta)$$

$$x|z = k \sim \mathrm{N}(\mu, \sigma\_k) \qquad \text{for } k \in \{0,1\}$$

GMM 모델의 모수는 $\theta, \mu, \sigma_0, \sigma_1$이며, 데이터 집합은 $D = \{x^{(i)}\}_{i=1}^{N}$으로 정의한다. Gaussian 분포 식은 3번 문제를 참고한다.

A. $x^{(i)}$가 생성된 Gaussian 분포를 알고 있다고 가정한다면 데이터 집합 D는 $D_{complete} = \{(z^{(i)}, x^{(i)})\}_{i=1}^{N}$와 같이 쓸 수 있으며 이를 완전한 데이터 집합(complete data)라고 부른다. 완전한 데이터 집합에 대한 log-likelihood $\log p(D_{complete})$을 GMM 모델의 모수들에 대한 식으로 쓰시오. $log\, p(D_{complete}) = log\, p \sum_{i=1}^{N} p\,(x^{(i)}, z^{(i)})$를 이용한다.

B. EM 알고리즘의 E 단계에서는 각 데이터 $x^{(i)}$에 대해 posterior 확률 $r^{(i)} = p(z^{(i)} = 1|x^{(i)})$를 구해야 한다. GMM 모델의 모수를 이용하여 $r^{(i)}$의 식을 쓰시오. Gaussian 분포는 $\mathrm{N}(x^{(i)}; \mu, \sigma)$로 나타낸다.

C. Complete data log-likelihood의 기대값에 대한 식을 쓰시오. EM 알고리즘의 M 단계에서는 이 값을 최대화하는 것을 목표로 한다. 기대값은 $r^{(i)}$와 Gaussian 분포 $\mathrm{N}(x^{(i)}; \mu, \sigma)$의 식으로 나타낸다.

D. M 단계에서 μ의 값을 구하는 식을 쓰시오. 이 단계에서는 σ_k는 고정된 값을 갖는다.
(힌트: 이전 문제에서 구한 complete data log-likelihood의 기대값을 μ에 대해 미분하여 최대값을 구한다.)

Considering only the parts of the objective which contain μ, we must maximize:

$$\sum_{i=1}^{N}(1-r^{(i)})\left(-\frac{(x^{(i)}-\mu)^2}{2\sigma_0^2}\right)+r^{(i)}\left(-\frac{(x^{(i)}-\mu)^2}{2\sigma_1^2}\right)$$

Differentiating with respect to μ and setting to zero:

$$0=\sum_{i=1}^{N}(1-r^{(i)})\left(\frac{x^{(i)}-\mu}{\sigma_0^2}\right)+r^{(i)}\left(\frac{x^{(i)}-\mu}{\sigma_1^2}\right)$$

After some rearranging, we get:

$$\mu=\frac{\frac{1}{\sigma_0^2}\sum_{i=1}^{N}(1-r^{(i)})x^{(i)}+\frac{1}{\sigma_1^2}\sum_{i=1}^{N}r^{(i)}x^{(i)}}{\frac{1}{\sigma_0^2}\sum_{i=1}^{N}(1-r^{(i)})+\frac{1}{\sigma_1^2}\sum_{i=1}^{N}r^{(i)}}$$

E. M 단계에서 σ_1의 값을 구하는 식을 쓰시오. 이 단계에서는 μ는 고정된 값을 갖는다.

Considering only the parts of the objective which contain σ_1, we must maximize:

$$\sum_{i=1}^{N}r^{(i)}\left[-\log(\sigma_1)+\left(-\frac{(x^{(i)}-\mu)^2}{2\sigma_1^2}\right)\right]$$

Differentiating with respect to σ_1 and setting to 0:

$$0=\sum_{i=1}^{N}r^{(i)}\left[-\frac{1}{\sigma_1}+\left(\frac{(x^{(i)}-\mu)^2}{\sigma_1^3}\right)\right]$$

After some rearranging, we get:

$$\sigma_1^2=\frac{\sum_{i=1}^{N}r^{(i)}(x^{(i)}-\mu)^2}{\sum_{i=1}^{N}r^{(i)}}$$

5. 입력, 은닉 레이어(hidden layer), 출력의 3개의 레이어로 구성된 신경망을 생각해 보자. 이 신경망은 입력 레이어에 데이터 $X = \{x_i\}$를 입력 받는다. 입력 레이어와 은닉 레이어 사이의 가중치 행렬은 $W = \{w_{i,j}\}$로 나타내며 $w_{i,j}$은 입력 레이어의 i번째 유닛의 출력이 은닉 레이어의 j번째 유닛으로 입력될 때 가중치를 의미한다. 출력 레이어의 유닛은 하나이고 은닉 레이어와 출력 레이어 사이의 가중치 벡터는 $V = \{v_i\}$로 나타내며, v_i는 은닉 레이어의 i번째 유닛의 출력이 출력 레이어에 입력될 때의 가중치이다. 은닉 레이어와 출력 레이어의 유닛들은 모두 Sigmoid 활성화 함수 $f(x) = \frac{1}{1+e^{-x}}$를 사용한다.

   A. 출력 레이어의 유닛의 출력값을 y라고 할 때, y를 f, X, W, V의 식으로 나타내시오.

   B. 신경망을 학습하기 위해 제곱 오차 $\ell = \frac{1}{2}(y-t)^2$을 손실 함수로 사용한다. 은닉 레이어의 유닛들의 출력들은 벡터 $H = \{h_i\}$로 놓는다. $w_{i,j}$의 gradient $\frac{\partial \ell}{\partial w_{i,j}}$를 y, x_i, $w_{i,j}$, v_i, h_i, f로 나타내시오.

6. PCA에서 최적의 부공간(subspace)을 얻는 과정은 데이터들의 공분산 행렬 $\Sigma = Cov(x)$의 eigen-decomposition을 통해 이루어진다. 공분산 행렬 Σ의 spectral decomposition은 $\Sigma = Q\Lambda Q^T$과 같다. Q는 직교 행렬, Λ는 대각 행렬을 나타낸다. Eigenvalue들은 내림차순으로 정렬되어 있고, 모두 다른 값들을 가진다고 가정한다.

   A. Eigen-decomposition을 이용해 Q와 Λ을 계산했다면 최적의 PCA 부공간을 구성하는 직교 기저 벡터 U는 어떻게 구할 수 있는지 한두문장으로 간단하게 설명하시오.

   B. 데이터 포인트 x의 PCA 부공간에서의 code 벡터는 $z = U^T(x-\mu)$로 주어진다. z의 각 차원들은 서로 독립임을 증명하시오. (힌트: $Cov(z) = U^T Cov(x) U$ 에서 시작한다. Σ의 spectral decomposition과 $U^T Q = (I \;\; 0)$을 이용한다)