기계학습 기말고사 (2020년도 2학기)

2020년 12월 16일 11:00-13:00

학번: 이름:

---

1. (6점) 아래의 문장들 중 맞는 것들을 고르시오.
   1) 은닉 레이어들이 없고 출력 레이어에서 logistic activation function를 사용하지 않는 신경망은 선형 결정 평면(linear decision boundary)을 출력하게 된다. ( )
   2) Gaussian Naïve Bayes 분류기는 임의의 클래스에 대해 입력 특징들의 공분산 행렬은 대각 행렬을 가정한다. ( )
   3) Gaussian Discriminant Analysis는 완전한 공분산 행렬(full covariance matrix)를 사용할 때 이차식 형태의 결정 평면(quadratic decision bound)를 가진다. ( )

2. (3점) 많은 기계학습 알고리즘들은 normalization이라고 부르는 학습 데이터 전처리 방법을 사용한다. 이는 벡터 형태의 학습데이터에 대해 데이터들의 각 차원 별로 데이터들의 평균을 0이 되도록 하고 분산이 1로 만드는 것이다. 주어진 학습 데이터의 j번째 차원의 평균이 μ_j , 분산이 σ_j일 때, 원래 데이터 x_j의 normalize된 값 $\hat{x}_j$에 대한 식을 쓰시오.

3. (4점) K-means 알고리즘은 두 단계에 걸쳐서 각 단계마다 다른 objective function을 사용한다. 이들을 각각 한 문장으로 설명하시오.

4. (3점) 아래와 같은 두 가지 종류의 선형 분류 모델들이 주어져 있다.

Model 1:

$$y = \mathbf{w}^\top \mathbf{x} + b$$
$$\mathcal{L}_{\mathrm{SE}}(y,t) = \tfrac{1}{2}(y-t)^2$$

Model 2:

$$z = \mathbf{w}^\top \mathbf{x} + b$$
$$y = \sigma(z)$$
$$\mathcal{L}_{\mathrm{SE}}(y,t) = \tfrac{1}{2}(y-t)^2$$

σ는 logistic function을 의미한다. Target t는 0 또는 1의 값을 갖는다. Model 2가 Model 1보다 더 좋은 모델이라고 말할 수 있는데 그 이유를 간단하게 설명하시오.

5. (6점) Support Vector Machine (SVM)은 아래와 같이 hinge loss를 최소화한다.

$$\min_{\mathbf{w},b} \sum_{i=1}^{N} \mathcal{L}_{\mathrm{H}}(y,t) + \frac{1}{2\gamma}\|\mathbf{w}\|_2^2$$

여기서 hinge loss는 아래와 같이 정의된다.

$$\mathcal{L}_{\mathrm{H}}(y,t) = \max(0, 1-ty)$$

1) Hinge loss의 총합이 0인 경우에는 (즉 $\sum_{i=1}^{N} L_H(y,t) = 0$) 모든 학습 데이터들은 정확하게 분류되어야 하는가? YES 또는 NO로 대답하고 그 이유도 쓰시오 (YES, NO만 표시하면 오답으로 간주)

2) Hinge loss를 아래와 같이 변경했다고 해보자. 다른 부분들은 위의 식과 동일하다. 이 경우 SVM은 제대로 학습되지 않는다. 그 이유는 무엇인지 설명하시오.

$$\mathcal{L}(y,t) = \max(0, -ty)$$

6. (9점) 다중레이어 퍼셉트론(MLP)의 레이어 하나가 아래와 같이 주어져 있다.

$$z_i = \sum_i w_{ij} h_j + b_i$$
$$y_i = \phi(z_i)$$

ϕ는 비선형 활성화 함수, h_j는 이 레이어의 입력(즉 이전 레이어의 출력), y_i는 이 레이어의 출력을 나타낸다. 역전파(backpropagation) 규칙을 이용해 $\bar{z}_i$, $\bar{h}_j$, $\bar{w}_{ij}$를 $\bar{y}_i$로 나타내시오. 퍼셉트론 모델의 손실 함수를 L로 놓으면 $\bar{a} = \frac{\partial L}{\partial a}$로 정의된다. ϕ의 미분은 ϕ'로 표시하시오.

$\bar{z}_i =$

$\bar{h}_j =$

$\bar{w}_{ij} =$

7. (6점) 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.

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

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.

1) (2점) 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)

2) (2점) 위에서 기술한 목적 함수에 대해 Laplace prior를 이용한 MAP의 해는 Gaussian prior를 이용한 MAP의 해와 어떻게 다른지 한 문장으로 설명하시오.

8. (6 점) Naïve Bayes classifier 는 이메일에 포함된 단어들(일반적으로 token 이라고 부름)의 상관 관계로부터 베이즈 규칙을 이용하여 스팸 메일일 확률을 계산한다. 스팸일 가능성이 높은 이메일들은 "대출"이라는 단어를 포함하고 있다고 가정해 보자. 이러한 메일들은 주로 대부업체들이 고금리의 대출을 홍보하는 내용들이어서, 이메일을 자주 사용하는 사람들은 이러한 메일들이 대부분 스팸이라는 것을 알고 있다. 반면 스팸 분류기는 이러한 내용을 알지 못하고 메일 속에 포함된 단어들을 이용하여 확률을 계산할 뿐이다. 즉 각 메일을 하나의 입력 데이터로 간주하고 이를 메일에 포함된 단어들로 구성된 특징 벡터로 사용하는 것이다.

   1) "대출"이라는 단어가 포함된 이메일이 스팸일 확률을 $P(S|W)$라고 하자. 아래에 제시된 확률들을 이용하여 $P(S|W)$를 나타내시오.

- $P(S)$: 임의의 메일이 스팸일 확률 (즉 전체 메일 중 스팸 메일의 비율)
- $P(W|S)$: 스팸 메일이 "대출"이라는 단어를 포함하고 있을 확률
- $P(H)$: 임의의 메일이 정상 메일일 확률 (즉 스팸 메일이 아닐 확률)
- $P(W|H)$: 정상 메일이 "대출"이라는 단어를 포함하고 있을 확률

   2) Naïve Bayes 는 입력 데이터의 feature 들이 "Independent and identically distributed random variable (I.I.D)"라고 가정한다. 그러나 실제로 스팸 메일에 들어 있는 단어들의 관계에서는 I.I.D 가 성립하지 않는다. 그 이유를 3 가지 이상 각각 한두문장으로 간략하게 기술하시오.

9. (4점) X가 discrete이거나 continuous이거나 상관없이, Naïve Bayes는 동일한 조건부 독립 가정을 적용해서 분류 문제에 사용할 수 있다. 이번 문제는 X값의 두가지 경우에 대해 Maximum Likelihood Estimation을 이용해서 Naïve Bayes를 푸는 방법에 관한 것이다.

   n개의 이진 변수로 구성된 확률 변수 X를 가정하자. 즉 $X =< X_1, X_2, \dots, X_n >$이고 X_i는 X의 i번째 attribute를 나타낸다. 첫번째 attribute(차원 또는 feature)인 X_1에 관련된 parameter들을 예측하는 것만 생각해 보자. Y는 K개의 가능한 값(y_k)들을 가질 수 있는 1차원 discrete 변수라고 가정하면, $P(X_1|Y = y_k)$은 베르누이 분포를 이용해서 아래와 같이 기술할 수 있다.

$$P(X_1 = x_{1j}|Y = y_k) = (\theta_{1k})^{x_{1j}}(1 - \theta_{1k})^{1-x_{1j}}$$

   위의 식에서 $j = 1 \dots M$은 M개의 전체 학습데이터에서 j번째 학습데이터를 의미하고, x_{1j}은 j번째 학습데이터의 첫번째 attribute를 의미한다. M개의 학습데이터들은 확률적으로 독립이고 동일 분포를 가진다고, 즉 IID를 만족한다고 가정한다. $\theta_{ij} = P(X_i|Y = y_j)$를 나타낸다고 가정하자.

   Maximum Likelihood Estimation(MLE)를 이용해서 θ_{1k}의 값을 계산하시오.

(힌트: j번째 학습데이터에 대한 likelihood를 $P(X_{1j} = x_{1j}|\theta_{1k}) = (\theta_{1k})^{x_{1j}}(1 - \theta_{1k})^{1-x_{1j}}$로 놓으면 전체 학습데이터에 대한 likelihood는 $L(\theta_{1k}) = \prod_{j=1}^{M} P(X_{1j}|\theta_{1k})^{I(Y^j=y_k)}$, $I(Y^j = y_k) = \begin{cases} 1 \text{ if } Y^j = y_k \\ 0 \text{ otherwise} \end{cases}$ 로 놓을 수 있다. $L(\theta_{1k})$를 미분해서 최소값을 구하는 방법을 이용한다.)