# 기계학습 기말고사 (2018년 2학기)

학번: 이름:

---

1. Quick questions (20점)

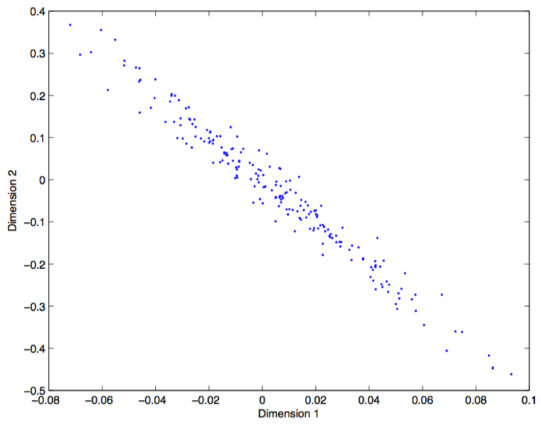
(a) 아래 그림과 같은 dataset을 이용하여 Gaussian Naïve Bayes(GNB)을 이용한 이진 분류기를 학습하려고 한다. Dataset은 GNB의 가정을 만족한다. 같은 클래스에 속하는 데이터들의 분산은 같다고 가정하자. 이때 아래의 경우들에서 decision boundary들은 어떻게 다른지 간단히 설명하시오.

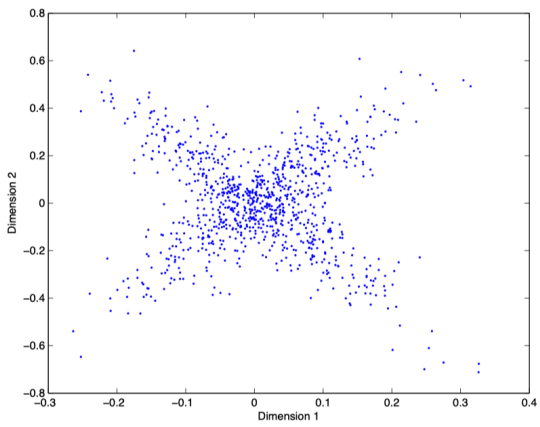
두 클래스의 분산이 같을 경우 vs. 두 클래스의 분산이 다른 경우

(b) 데이터 D와 hypothesis H (예측하고자 하는 값, 클래스 등등)에 대해 아래의 식들이 참인지 거짓인지 쓰시오.

A. $\sum_h P(H = h|D = d) = 1$
B. $\sum_h P(D = d|H = h) = 1$
C. $\sum_h P(D = d|H = h)P(H = h) = 1$

(c) 주성분 분석(Principal Component Analysis, PCA)은 데이터의 차원을 축소시켜 표현하기 위해 가장 변화가 많이 일어나는 방향(즉 주성분)으로 학습 데이터를 사영(projection)시키는 방법이다. 아래의 그림과 같이 두 종류의 2 차원 데이터 세트가 주어져 있다. 각각의 데이터 세트에 대해 첫번째, 두번째 주성분 축을 그림에 표시하시오. (데이터 차원 축소는 고려하지 않음)



2. Expectation and Maximization (EM) (10점)

이진 확률 변수 $X2$의 값이 다른 이진 확률변수 $X1$의 값에 대해 조건부로 일어나는 아래의 dataset이 주어져 있다.

| Example | $X1$ | $X2$ |
|---|---|---|
| 1. | 0 | 1 |
| 2. | 0 | 0 |
| 3. | 1 | 0 |
| 4. | 1 | ? |
| 5. | 0 | 1 |

Dataset의 example 4에서는 X2의 값이 누락되어 있는데 이를 EM 알고리즘을 이용하려 풀려고 한다. 아래의 3개의 latent parameter들($\hat{\theta}_{X1=1}$, $\hat{\theta}_{X2=1|X1=1}$, $\hat{\theta}_{X2=1|X1=0}$)을 구하려고 하며, E 단계와 M 단계를 몇 번 수행하여 예측한 parameter들의 값은 아래와 같다.

$\hat{\theta}_{X1=1} = \hat{P}(X1 = 1) = 0.4$

$\hat{\theta}_{X2=1|X1=1} = \hat{P}(X2 = 1|X1 = 1) = 0.4$

$\hat{\theta}_{X2=1|X1=0} = \hat{P}(X2 = 1|X1 = 0) = 0.66$

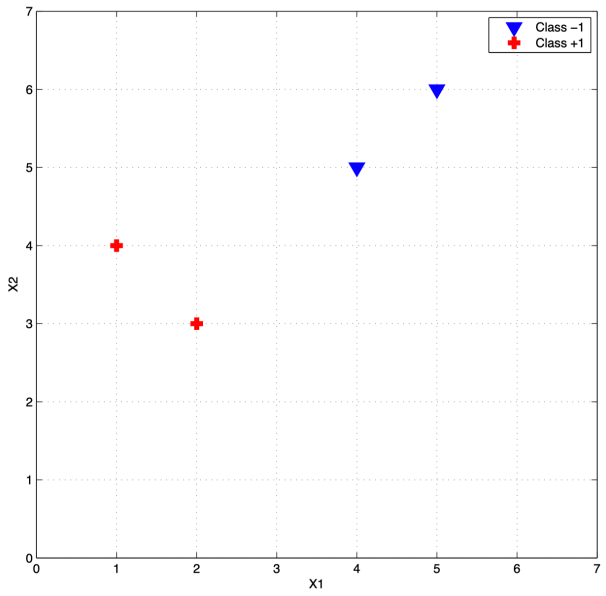
(a) 다음 E 단계에서는 example 4에서 X2의 기대값을 구할 수 있다. 계산하시오. (힌트: $\hat{P}(X2 = 1|X1 = 1; \hat{\theta})$을 구한다)

(b) 다음 M 단계에서 각 latent parameter들의 값들이 어떻게 되는지 계산하시오.

3. Support Vector Machines (15점)

SVM은 두 클래스로부터 가장 큰 margin을 가지는 decision boundary를 학습한다. 아래의 그림처럼 4개의 점들로 구성되어 있는 작은 데이터세트로부터 SVM을 학습한다고 가정하자. 데이터세트는 그림에서 십자 모양과 역삼각형의 모양의 두 개의 클래스들로 구성되어 있으며, 각 클래스들의 레이블은 -1와 1이라고 하자.

(a) Decision boundary 직선을 나타내는 가중치 벡터 $W$와 바이어스 $b$를 계산하시오.




(b) 위의 그림에 decision boundary에 직선을 그리고, support vector에 해당하는 데이터들을 동그라미로 표시하시오.

4. Naïve Bayes classifier 를 이용한 스팸 메일 분류기 설계 (10 점)
   Naïve Bayes classifier 는 이메일에 포함된 단어들(일반적으로 token 이라고 부름)의 상관관계로부터 베이즈 룰을 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 feature vector 로 사용하는 것이다.

   (a) "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
   - $P(H)$ : 임의이 메일이 정상 메일일 확률 (즉 스팸 메일이 아닐 확률)
   - $P(W|H)$ : 정상 메일이 "대출"이라는 단어를 포함하고 있을 확률

   (b) 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.

5. Naïve Bayes (15점)
   X가 discrete이거나 continuous이거나 상관없이, Naïve Bayes는 동일한 조건부 독립 가정을 적용해서 분류 문제에 사용할 수 있다. 이번 문제는 X값의 두가지 경우에 대해 Maximum Likelihood Estimation 을 이용해서 Naïve Bayes를 푸는 방법에 관한 것이다.
   (a) n개의 이진 변수로 구성된 확률 변수 $X$를 가정하자. 즉 $X =< X_1, X_2, \dots, X_n >$이고 $X_i$는 $X$의 i번째 attribute를 나타낸다. 첫번째 attribute(차원 또는 feature)인 $X_1$에 관련된 parameter들을 예측하는 것만 생각해 보자. $Y$는 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 $j = 1 \ldots M$은 M개의 전체 학습데이터에서 j번째 학습데이터를 의미하고, $x_{1j}$은 j번째 학습데이터의 첫번째 attribute를 의미한다. M개의 학습데이터들은 확률적으로 독립이고 동일분포를 가진다고, 즉 IID를 만족한다고 가정한다. $\theta_{ij} = P(X_i|Y = y_j)$를 나타낸다고 가정하자.

Maximum Likelihood Estimation(MLE)를 이용해서 $\theta_{1k}$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

$$L(\theta_{1k}) = \prod_{j=1}^{M} P(X_{1j}|\theta_{1k})^{I(Y^j = y_k)},\ I(Y^j = y_k) = 1 \text{ if } Y^j = y_k;\ I(Y^j = y_k) = 0 \text{ otherwise}$$

로 놓을 수 있다. $L(\theta_{1k})$를 미분해서 최소값을 구하는 방법을 이용한다.

(b) 이번에는 $X_i$가 가우시안 분포를 가진 continuous 확률 변수라고 가정하자. 즉

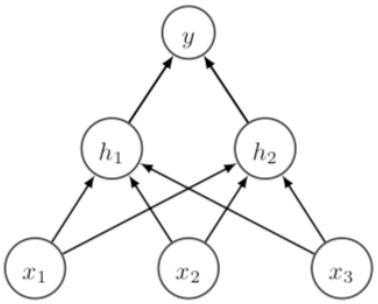
$$P(X_i = x_{ij} | Y = y_k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} \exp\left(-\frac{(x_{ij} - \mu_{ik})^2}{2\sigma_{ik}^2}\right)$$

가우시안 분포의 분산은 클래스 변수 $Y$와 attribute $X_i$에 독립이라고 가정하자. 그러므로 $\sigma_{ik} = \sigma$로 놓을 수 있다. MLE를 사용하여 가우시안 분포의 평균 $\mu_{ik}$를 구하시오.

**힌트:** 학습데이터 전체의 likelihood $L(\mu_{ik}; \sigma) = \prod_{j=1}^{M} \left[\frac{1}{\sigma_{ik}\sqrt{2\pi}} \exp\left(-\frac{(x_{ij}-\mu_{ik})^2}{2\sigma_{ik}^2}\right)\right]^{I(Y^j = y_k)}$ 를 미분하여 최소값일때의 $\mu_{ik}$를 구한다. $I(Y^j = y_k)$는 앞의 문제와 동일하다.

6. [Gaussian mixture models] (10점)

Gaussian mixture models(GMM)에서의 E-step은 각 data point들이 어떤 Gaussian model로부터 생성되었는지 예측하는 확률 모델 $Z: P(Z_j|X_i), i = 1, \dots, n, j = 1, \dots, k$ 를 구한다. GMM은 2개의 Gaussian 분포 $z_1$와 $z_2$로 구성되어 있으며 각각 평균 $\mu_1$, $\mu_2$와 같은 분산 1을 갖는다. 데이터들의 아래와 같은 $z_1$와 $z_2$로 구성된 prior 분포를 갖는다고 가정한다.




$$N(\mu_1, 1) \times 0.5 + N(\mu_2, 1) \times 0.5$$

$x_1 = 2$인 데이터와 $\mu_1 = 2$, $\mu_2 = 1$를 가정하자. 데이터 $x_1$이 다음 iteration의 E-step에서 각 Gaussian 분포에 속할 확률 $p(z_1|x_1)$과 $(z_2|x_1)$를 구하시오.

**힌트:** 확률변수 $y \sim N(0,1)$에 대해 $p(y = 0) = 0.4$, $p(y = 0.5) = 0.35$, $p(y = 1) = 0.24$, $p(y = 1.5) = 0.24$, $p(y = 1.5) = 0.13$, $p(y = 2) = 0.05$ 의 값을 가짐

7. Neural Networks (20점)

아래의 그래프는 한개의 hidden layer를 가진 간단한 신경망 구조를 나타낸다. Input layer에서는 3차원 데이터 $x = (x_1, x_2, x_3)$를 사용한다. Hidden layer는 2개의 unit $h = (h_1, h_2)$로 구성되어 있다. Output layer는 하나의 unit $y$를 사용한다. 계산의 편의를 위해 bias는 사용하지 않는다.

Hidden layer와 output layer의 activation function으로는 linear rectified unit $\sigma(z) = \max(0, z)$ 를 사용한다. Loss function으로는 $l(y, t) = \frac{1}{2}(y - t)^2$를 사용한다. $t$는 target value이고 $y$는 output unit을 의미한다. $W$를 input layer와 hidden layer를 연결하는 weight matrix로, $V$를 hidden layer와 output layer를 연결하는 weight matrix로 놓는다. 각 변수들은 아래와 같이 초기화된다.

$$W = \begin{bmatrix} 1 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix}, \; V = [0 \quad 1], \; x = [1,2,1], \; t = 1$$

학습 데이터는 위의 값을 $(x, t)$를 포함해 최소 한 개 이상을 가지고 있다고 가정한다.

(a) $x \rightarrow y$의 함수식을 $\sigma$, W, V를 이용해 쓰시오. 즉 $y = f(x)$의 형태에 해당하는 $f$를 $\sigma$, W, V로 나타내는 것을 생각해 보면 됨.

(b) $x = (1,2,1)$가 입력으로 주어져 있고 target는 $t = 1$이라고 가정하자. Output $y$의 값을 계산하시오. Matrix 곱셈을 이용해서 계산하시오.

(c) weight들에 대한 loss function의 gradient를 계산하시오. 즉 다음의 값들을 계산하는 것이다.

- ✓ $V$에 대한 loss function의 gradient, $\frac{\partial l}{\partial V}$
- ✓ $W$에 대한 loss function의 gradient, $\frac{\partial l}{\partial W}$
- ✓ 문제 앞 부분에서 주어진 W, $V, x, y$의 값들을 이용해 gradient들의 값을 계산할 것