CSC 411 Fall 2018
Machine Learning and Data Mining

Homework 4

Family name: Bhatia Given name: Pooja

Solution 1

(a)

Input:

$W \times H \times C$ Kernal size K

Output Maps M

Units = pixels x output units

$= WH \times M = WHM$

Weights = size of each piece of input x No of filters

$= K^2C \times M$

$= K^2CM$

Connections = Values processed by a kernal X output units

$= K^2C \times WHM$

$= K^2C\ WHM$.

Layer 1.

M = 96 Kernals → $\overset{K}{11} \times 11 \times 3$ each

stride = 4

Input = $\overset{W}{224} \times \overset{H}{224} \times \overset{C}{3}$

New Iput: $W = H = 55$ $C = 3$

units = $55 \times 55 \times 96 = \boxed{290,400}$

weight = $K^2CM = 11^2 \times 3 \times 96 = \boxed{34,848}$

Connections = $K^2CMWH = 11^2 \times 3 \times 96 \times 55 \times 55$

$= \boxed{105,415,200}$

Layer 2.

M = 256   size → $5 \times 5 \times 48 = K^2C$

$W = H = 55/2 = 27$ (max Pooling)

units = $WHM = 27^2 \times 256 = \boxed{186,624}$

weight = $K^2CM = 5^2 \times 48 \times 256 = \boxed{307,200}$

Connections = $K^2C\,WHM = 5^2 \times 48 \times 27^2 \times 256$

$= \boxed{223,948,800}$

Layer 3

$M = 384 \quad K^2C = 3 \times 3 \times 256$

$W = H = 27/2 = 13$ (max pooling)

Units $= WHM = 13^2 \times 384 = \boxed{64{,}896}$

Weights $= K^2CM = 3 \times 3 \times 256 \times 384$

$= \boxed{884{,}736}$

Connections $= K^2C\,WHM = 3^2 \times 256 \times 13^2 \times 384$

$= \boxed{149{,}520{,}384}$

Layer 4

$M = 384 \quad K^2C = 3 \times 3 \times 192$

units $= 13^2 \times 384 = \boxed{64{,}896}$

weight $= 3^2 \times 192 \times 384 = \boxed{663{,}552}$

Connection $= 3^2 \times 192 \times 13^2 \times 384$

$= \boxed{112{,}140{,}288}$

layer 5

$M = 256 \qquad K^2C = 3 \times 3 \times 192$

units $= 13^2 \times 256 = \boxed{43,264}$

weight $= 3^2 \times 192 \times 256 = \boxed{442,368}$

connection
$= 3^2 \times 192 \times 13^2 \times 256$
$= \boxed{74,760,192}$

layer 6

units $= 4096$
weight $= 6^2 \times 256 \times 4096 = \boxed{37,748,736}$
connection $= 6^2 \times 256 \times 4096 = \boxed{37,748,736}$

layer 7

units $= 4096$
weight $= 4096 \times 4096 = \boxed{16,777,216}$
connection $= 4096^2 = \boxed{16,777,216}$

layer 8

Units = $\boxed{1000}$

weight = 4096 x 1000 = $\boxed{4,096,000}$

Connection = 4096 x 1000 = $\boxed{4,096x000}$

layer	#Units	#weights	#Connections
convolution Layer 1	290,400	34,848	105,415,200
convolution layer 2	186,624	307,200	223,948,800
convolution layer 3	64,896	884,736	149,520,384
convolution layer 4	64,896	663,552	112,140,288
Convolution layer 5	43,264	442,368	74,760,192
Fully connected layer 1	4096	37,748,736	37,748,736
Fully connected Layer 2	4096	16,777,216	16,777,216
Output Layer	1000	4,096,000	4,096,000

Solution 1.
(b)

(i)

Fully connected layers contribute to the most number of paramotrs i.e the weights.

weights = $K^2 C M$
(hence reducing K and m would help)
Also, Alexnet works with same accuracy even if its parameters are reduced certain times.

(b)
(ii) Convolution Layers contributes to most number of connections

connection = $K^2 C M W H$

hence, suggestion: Reduce the No. of Kernals; or reduce size of kernals. would both help.

soln Q2. Gaussian Naive Bayes

(a) X is (x_1 . . . . x_d) i.e d feature X.

$X = (X_1, X_2 \ldots\ldots X_d)$, where each X_i is a continous random valiable.

$Y \in (1, 2 \ldots\ldots K)$ but to start and for simplicity let's first Assume that Y is Boolean and has parameter say $\pi = P(Y=1)$

Given, that for each X_i, $P(X_i \mid Y = \alpha_K)$ is a Gassian Distribution of the form $N(\mu_{iK}, \sigma_i)$

For all i and $j \neq i$, X_i and X_j are conditionally independent given Y.

σ_i varies from attribute to attribute but doesn't depend on Y.

NOW,

$$P(Y=1 \mid X) = \frac{P(Y=1)\,P(X \mid Y=1)}{P(Y=1)\,P(X \mid Y=1) + P(Y=0)\,P(X \mid Y=0)}$$

Divide both numerator and denominator by numerator.

$$p(Y=1|X) = \frac{1}{1 + \frac{p(Y=0)\,p(X|Y=0)}{p(Y=1)\,p(X|Y=1)}}$$

or equivalently,

$$p(Y=1|X) = \frac{1}{1 + \exp\left(\ln\left(\frac{p(Y=0)\,p(X|Y=0)}{p(Y=1)\,p(X|Y=1)}\right)\right)}$$

As taken conditional independence assumption, we can say

$$p(Y=1|X) = \frac{1}{1 + \exp\left(\ln\frac{p(Y=0)}{p(Y=1)} + \sum_i \ln\frac{p(X_i|Y=0)}{p(X_i|Y=1)}\right)}$$

$$= \frac{1}{1 + \exp\left(\ln\frac{1-\pi}{\pi} + \sum_i \ln\frac{P(X_i|Y=0)}{P(X_i|Y=1)}\right)}$$

$\rightarrow$ 1

Since it is given that $P(X_i|Y=\alpha_K)$ Is Gaussian, hence

$$\sum_i \ln \frac{P(X_i|Y=0)}{P(X_i|Y=1)} = \sum_i \ln \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(X_i-\mu_{i0})^2}{2\sigma_i^2}\right)$$

$$= \sum_i \ln \exp\left(\frac{(X_i-\mu_{i1})^2-(X_i-\mu_{i0})^2}{2\sigma_i^2}\right)$$

$$= \sum_i \left(\frac{(X_i-\mu_{i1})^2-(X_i-\mu_{i0})^2}{2\sigma_i^2}\right)$$

$$= \sum_i \left(\frac{(X_i^2-2X_i\mu_{i1}+\mu_{i1}^2)-(X_i^2-2X_i\mu_{i0}+\mu_{i0}^2)}{2\sigma_i^2}\right)$$

$$= \sum_i \left(\frac{2X_i(\mu_{i0}-\mu_{i1})+\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}\right)$$

$$= \sum_i \left(\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}X_i + \frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}\right)$$

$\rightarrow$ 2

Now, this expression is a linear weighted sum of x_i's. Substituting expression (2) in (1), we get

$$P(Y=1|X) = \frac{1}{1+\exp\left(\ln\frac{1-\pi}{\pi} + \sum_i \left(\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2} x_i + \frac{(\mu_{i1}^2-\mu_{i0}^2)}{2\sigma_i^2}\right)\right)}$$

$$\Rightarrow \quad P(Y=1|X) = \frac{1}{1+\exp\left(w_0 + \sum_{i=1}^{d} w_i x_i\right)}$$

$w_i \cdots\cdots w_d$ are given by

$$w_i = \frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}$$

and

$$w_0 = \ln\frac{1-\pi}{\pi} + \sum_i \frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}$$

Now for

$P(Y=0|X) = 1 - P(Y=1|X)$

$$P(Y=0|X) = 1 - P(Y=1|X) = \frac{\exp(w_0 + \sum_{i=1}^{d} w_i X_i)}{1+\exp(w_0 + \sum_{i=1}^{d} w_i X_i)}$$

we considered only cases where Y was a boolean variable, but NOW if Y can take any of discrete class laybel values i.e $y \in (1 \ldots k)$ that is y can take values $(y_1 \ldots y_K)$

Then

$P(Y=\alpha_K|X)$ for $Y=\alpha_1$ $Y=\alpha_2 \ldots$ $Y=\alpha_{K-1}$ is:

$$P(Y=\alpha_K|X) = \frac{\exp(w_{K0} + \sum_{i=1}^{d} w_{Ki} X_i)}{1+\sum_{j=1}^{K-1} \exp(w_{j0} + \sum_{i=1}^{n} w_{ji} X_i)}$$

$\therefore$

$$\boxed{P(y=k|X,\mu,\sigma) = \frac{\exp(w_{K0} + \sum_{i=1}^{d} w_{Ki} X_i)}{1+\sum_{j=1}^{K-1} \exp(w_{j0} + \sum_{i=1}^{n} w_{ji} X_i)}}$$

Ans

Ans (2)
b

$$l(\theta; D) = -\log P(y^{(1)}, x^{(1)}, y^{(2)}, x^{(2)} \dots y^{(N)} x^{(N)} \mid \theta)$$

Assuming data are i.i.d

$$l(\theta; D) = -\log\left(\prod_{i=1}^{N} P(y^{(i)}, x^{(i)} \mid \theta)\right)$$

we can take Σ as i.i.d data.

$$= -\sum_{i=1}^{N} \log P(y^{(i)}, x^{(i)} \mid \theta)$$

$$= -\Sigma_N \log P(x^{(i)}, y^{(i)} \mid \theta) \rightarrow (1)$$

Since we can write

$$P(x, y \mid \theta) = \log P(y \mid \theta) + \log P(x \mid y, \theta)$$

putting it in equation (1)

$$l(\theta; D) = -\sum_N \log P(x^{(i)}, y^{(i)} | \theta)$$

$$\boxed{l(\theta; D) = -\sum_N \left(\log P(y^{(i)} | \theta) + \log P(x^{(i)} | y^{(i)}, \theta)\right)}$$

Ans

Q2 c. using the Multivarate Gassian Distribution Equation since in the question it mentions shared variance (σ_i^2).

$$N(X \mid \mu, \Sigma) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(X-\mu)^T \Sigma^{-1} (X-\mu)\right\} \rightarrow (1)$$

where

X is D-dimensional vector

μ is D dimensional mean vector

Σ is D x D covariance matrix with determinant $|\Sigma|$

Σ (covariance matrix) is a matrix whose (i, j) entry is covariance i.e

$$\Sigma_{ij} = \text{cov}(x_i, x_j)$$

$$= E[(x_i - \mu_i)(x_j - \mu_j)]$$

$$= E[(x_i x_j)] - \mu_i \mu_j$$

Diagonal entries are variance of

each elements, hence using equation ① in the question.

$\theta = [\mu, \Sigma, \alpha]$

$Z = \sqrt{(2\pi)^D |\Sigma|} = (2\pi)^{D/2} |\Sigma|^{1/2}$

$p(x|y) = \frac{1}{Z} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$

$\log L(\theta) = \log p(x, y|\theta) = \log p(y|\theta) + \log p(x|y,\theta)$

$= \sum_{i=1}^{N} \log \alpha_{y^{(i)}} - \log Z - \frac{1}{2}(x^{(i)} - \mu_{y^{(i)}})^T \Sigma_{y^{(i)}}^{-1} (x^{(i)} - \mu_{y^{(i)}})$ → ①

since we are aiming at maximum likelihood that is $\arg\max_\theta \log L(\theta)$

this can happen when

$\Sigma_K \alpha_K = 1$

Taking derivative w.r.t μ

$$\frac{\partial \log L}{\partial \mu_K} = -\sum_{i=0}^{N} \mathbb{1}(y^{(i)} = K)\Sigma^{-1}(x^{(i)} - \mu_K) = 0$$

$\therefore$

$$\boxed{\mu_K = \frac{\sum_{i=1}^{N} \mathbb{1}(y^{(i)} = K)\, x^{(i)}}{\sum_{i=1}^{N} \mathbb{1}(y^{(i)} = K)}}$$ Ans

Now take derivative of equation (1) w.r.t Σ^{-1}

$$\frac{\partial \log L}{\partial \Sigma_K^{-1}} = -\sum_{i=0}^{N} \mathbb{1}(y^{(i)} = K)\left[-\frac{\partial \log Z_K}{\partial \Sigma_K^{-1}} - \frac{1}{2}(x^{(i)} - \mu_K)(x^{(i)} - \mu_K)^T\right] = 0 \quad \to (2)$$

$$Z = (2\pi)^{D/2} |\Sigma_K|^{1/2}$$

$$\frac{\partial \log Z_K}{\partial \Sigma_K^{-1}} = \frac{1}{Z_K} \frac{\partial Z_K}{\partial \Sigma_K^{-1}}$$

$$= (2\pi)^{-D/2} |\Sigma_K|^{-1/2} (2\pi)^{D/2} \frac{\partial (|\Sigma_K^{-1}|)^{-1/2}}{\partial \Sigma_K^{-1}}$$

$$= |\Sigma_K^{-1}|^{1/2} (-\tfrac{1}{2}) |\Sigma_K^{-1}|^{3/2} |\Sigma_K^{-1}| \Sigma_K^{T} = -\frac{1}{2} \Sigma_K$$

(using property that $\Sigma = \Sigma^T$)

Now substituting this back in equation (2) we get

$$\frac{\partial \operatorname{Log} L}{\partial \Sigma_K^{-1}} = -\sum_{i=0}^{N} \mathbb{1}(y^{(i)} = K) \left[\frac{1}{2} \Sigma_K - \frac{1}{2} (x^{(i)} - \mu_K)(x^{(i)} - \mu_K)^T\right] = 0$$

$$\boxed{\Sigma_K = \frac{\sum_{i=1}^{N} \mathbb{1}(y^{(i)} = K)(x^{(i)} - \mu_K)(x^{(i)} - \mu_K)^T}{\sum_{i=1}^{N} \mathbb{1}(y^{(i)} = K)}}$$

Ans

Q(2)
(d)

As suggested in the question using Lagrange multiplier

$$\frac{\partial L(\theta)}{\partial \alpha_K} + \lambda \frac{\partial \Sigma_K \alpha_K}{\partial \alpha_K} = 0$$

Note: Since we need to maximize we are adding the λ term.

$$\lambda = -\sum_{i=1}^{N} \mathbb{1}(y^{(i)} = K)\frac{1}{\alpha_K}$$

$$\alpha_K = \frac{-\sum_{i=1}^{N} \mathbb{1}(y^{(i)} = K)}{\lambda} \rightarrow (1)$$

Since again we are wanting to maximize the likehood as that is possible when

$$\Sigma_K \alpha_K = 1 \Rightarrow \lambda = -N$$

Substituting value of $\lambda = -N$ in equation (1)

$$\alpha_K = \frac{-\sum_{i=1}^{N} \mathbb{1}(y^{(i)} = K)}{-N}$$

$$\boxed{\alpha_K = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(y^{(i)} = K)}$$

Ans.