ECE M146 Discussion 7
Introduction to Machine Learning
Instructor: Lara Dolecek
TA: Zehui (Alex) Chen

1. We are given a training set $\{(x^{(i)}, y^{(i)}); i = \{1, \cdots, m\}\}$, where $x^{(i)} \in R^n$ and $y^{(i)} \in \{0, 1\}$. We consider the Gaussian Discriminant Analysis (GDA) model, which models $P(x|y)$ using multivariate Gaussian. Writing out the model, we have:

$$P(y = 1) = \underline{\phi} = 1 - P(y = 0) \qquad \text{prior}$$

$$P(x|y = 0) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_0)^T \underline{\Sigma}^{-1}(x - \underline{\mu_0})\right)$$

$$P(x|y = 1) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \underline{\mu_1})\right)$$

The log-likelihood of the data is given by:

$$L(\phi, \mu_0, \mu_1, \Sigma) = \ln P(x^{(i)}, \cdots, x^{(m)}, y^{(i)}, \cdots, y^{(m)}) = \ln \prod_{i=1}^{m} P(x^{(i)}|y^{(i)})P(y^{(i)}).$$

In this exercise, suppose we already find μ_0 and μ_1, we want to maximize $L(\phi, \mu_0, \mu_1, \Sigma)$ with respect to Σ.

(a) Write down the explicit expression for $P(x^{(1)}, \cdots, x^{(m)}, y^{(1)}, \cdots, y^{(m)})$ and $L(\phi, \mu_0, \mu_1, \Sigma)$.

$$P(x^{(1)}, \ldots, x^{(m)}, y^{(1)}, \ldots y^{(m)}) = \prod_{i=1}^{m} P(x^{(i)}, y^{(i)})$$

$$= \prod_{i=1}^{m} \left[\frac{1-\phi}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu_0)^T \Sigma^{-1}(x-\mu_0)\right)\right]^{1-y^{(i)}}$$

$$\times \left[\frac{\phi}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1)\right)\right]^{y^{(i)}}$$

$$L(\phi, \mu_0, \mu_1, \Sigma)$$

$$= \sum_{i=1}^{m} \Big\{ (1-y^{(i)}) \left[\ln(1-\phi) - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma|) - \frac{1}{2}(x^{(i)}-\mu_0)^T \Sigma^{-1}(x^{(i)}-\mu_0)\right]$$

$$+ y^{(i)} \left[\ln(\phi) - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma|) - \frac{1}{2}(x^{(i)}-\mu_1)^T \Sigma^{-1}(x^{(i)}-\mu_1)\right]$$

(b) Differentiate $L(\phi, \mu_0, \mu_1, \Sigma)$ with respect to Σ and set it to 0. Show that the maximum likelihood result for Σ is:

$$\Sigma = \frac{1}{m} \sum_{i=1}^{m} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T.$$

Hints: You may use the following properties without proof: $a = Tr(a)$ for scalar a; $Tr(A) + Tr(B) = Tr(A + B)$; $\boxed{\frac{\partial \ln |A|}{\partial A} = A^{-T}}$; $\boxed{\frac{\partial Tr(A^{-1}B)}{\partial A} = -(A^{-1}BA^{-1})^T}$.

$$L(\phi, \mu_0, \mu_1, L) = -\frac{m}{2}\ln(|\Sigma|) - \frac{1}{2}\sum_{i=1}^{m} (x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}}) + \text{- - -}$$

$$= -\frac{m}{2}\ln(|\Sigma|) - \frac{1}{2}\sum_{i=1}^{m} Tr\left((x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}})\right)$$

$$= -\frac{m}{2}\ln(|\Sigma|) - \frac{1}{2}\sum_{i=1}^{m} Tr\left(\Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T\right)$$

$$= -\frac{m}{2}\ln(|\Sigma|) - \frac{m}{2} Tr\left(\Sigma^{-1} \underbrace{\frac{1}{m}\sum_{i=1}^{m} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T}_{S}\right)$$

$$= -\frac{m}{2}\ln(|\Sigma|) - \frac{m}{2} Tr(\Sigma^{-1} S)$$

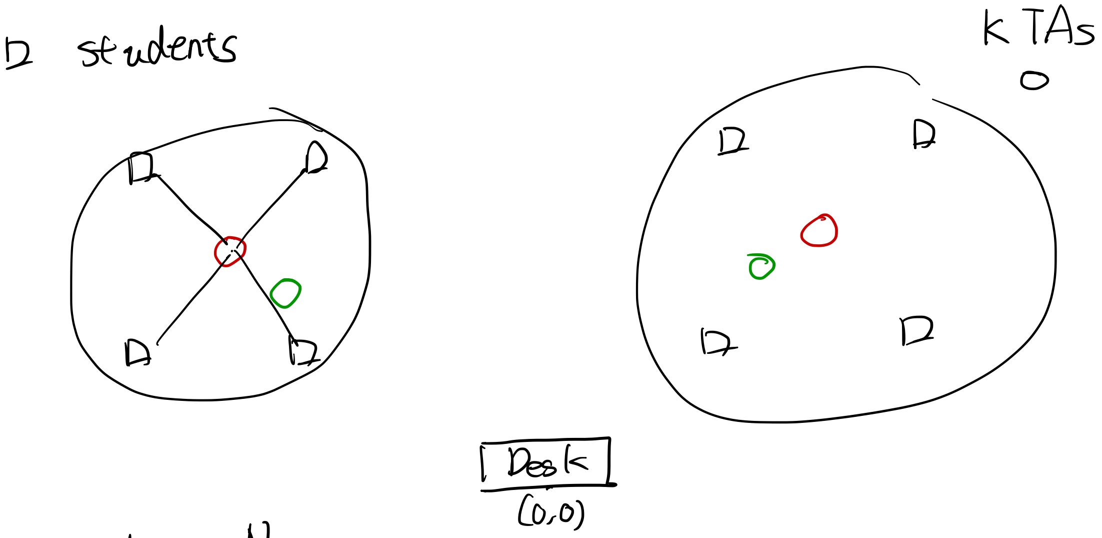
$$\frac{\partial L}{\partial \Sigma} = -\frac{m}{2}\Sigma^{-T} + \frac{m}{2}(\Sigma^{-1} S \Sigma^{-1})^T = 0$$

$$\Sigma^{-T} + \Sigma^{-T} S^T \Sigma^{-T} = 0$$

$$\Sigma^{-T} S^T = I$$

$$\Sigma = S$$

2. Here is an example where we would want to regularize clusters. Suppose n students are seated for taking an endterm exam in an $\mathbf{R}^2$ Euclidean room. There are K teaching assistants who must collect the answer scripts once the time is up. The TAs need to figure out good locations to position themselves so that the students can walk to the nearest TA and submit their answers. Once the TAs have all the answer sheets, they must return to the front desk located at (0,0) while handling the returned answer sheets carefully. To reduce the possibility of mishaps related to handling of the papers, write down an objective which can be used to minimize the total distance that both students and TAs need to walk to bring the papers to the front desk. Assume that everyone can walk by taking the shortest path between two points.



$$J = \sum_{k=1}^{K} \sum_{n=1}^{N} r_{nk} \|x_n - \mu_k\|_2^2$$

$$J = \sum_{k=1}^{K} \|\mu_k\|_2^2 + \sum_{n=1}^{N} r_{nk} \|x_n - \mu_k\|_2^2$$

3. We learned that the μ that ~~maximize~~ minimize

$$\sum_{i=1}^{N}(x_i - \mu)^2$$

for $x_i \in \mathbf{R}$ is given by the mean of $\{x_1, \cdots, x_N\}$,i.e., $\mu^* = \frac{1}{N}\sum_{i=1}^{N} x_i$.

Show that the μ that ~~maximize~~ minimize

$$J = \sum_{i=1}^{N}(x_i - \mu)^0$$

is given by the mode of $\{x_1, \cdots, x_N\}$. What if $x_i \in \mathbf{R}^n$?

$0^0 = 0$ and $a^0 = 1 \quad \forall a \neq 0$

If $\mu \notin \{x_1, \ldots, x_n\}$ $\quad J = N$

If $\mu \in \{x_1, \ldots x_n\}$ $\quad J = N - |\{x_i | x_i = \mu\}|$

To minimize J, we selest μ to be the data with maximum frequency.

Mode!

$$\frac{\partial \log \det X}{\partial X} = X^{-T} \qquad X \in R^{n\times n}$$

(i, j) minor of X : M_{ij} is the determinant of the $(n-1)\times(n-1)$ matrix after removing the i-th row and j-th column of X.

cofactor matrix of X : C is a matrix with $C_{ij} = (-1)^{i+j} M_{ij}$

adjugate matrix of X : $\operatorname{adj} X = C^T$

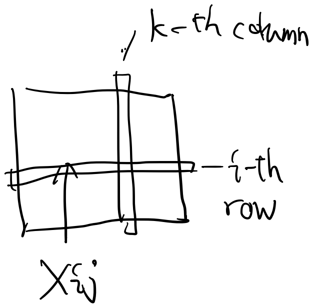
cofactor expansion of determinant :

$$\det X = \boxed{\sum_{k=1}^{n} X_{ik} C_{ik}}$$

Inverse : $X^{-1} = \frac{1}{\det X} \operatorname{adj} X \qquad X^{-T} = \boxed{\frac{1}{\det X} C}$

$$\frac{\partial (\det X)}{\partial X_{ij}} = \sum_{k=1}^{n} \left[\underbrace{\frac{\partial X_{ik}}{\partial X_{ij}} C_{ik}} + \underbrace{X_{ik} \frac{\partial C_{ik}}{\partial X_{ij}}} \right]$$

$$= C_{ij} + 0$$



$$\frac{\partial \log \det X}{\partial X_{ij}} = \frac{1}{\det X} \frac{\partial (\det X)}{\partial X_{ij}} = \boxed{\frac{1}{\det X} C_{ij}} = (X^{-T})_{ij}$$

Multivariat Gaussian.

$$f_{X,Y}(x,y) = \frac{1}{2\pi \sigma_X \sigma_Y \sqrt{1-\rho^2}} e^{\underbrace{-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_X^2} + \frac{y^2}{\sigma_Y^2} - \frac{2\rho x y}{\sigma_X \sigma_Y}\right)}_{t}}$$

a) Marginal of Y

$$f_Y(y) = \int_{x=-\infty}^{\infty} f_{X,Y}(x,y)\, dx$$

$$t = -\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\underline{\sigma_X^2}} + \frac{y^2}{\sigma_Y^2} - \frac{2\rho x y}{\underline{\sigma_X \sigma_Y}}\right)$$

$$= -\frac{1}{2(1-\rho^2)}\left[\left(\frac{x}{\sigma_X} - \frac{\rho y}{\sigma_Y}\right)^2 - \frac{\rho^2 y^2}{\sigma_Y^2} + \frac{y^2}{\sigma_Y^2}\right]$$

$$= -\frac{1}{2(1-\rho^2)\sigma_X^2}\left[x - \frac{\rho \sigma_X y}{\sigma_Y}\right]^2 - \frac{y^2}{2\sigma_Y^2}$$

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\,\sigma_Y} e^{-\frac{y^2}{2\sigma_Y^2}} \underbrace{\int_{x=-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma_X\sqrt{1-\rho^2}} e^{-\frac{\left(x - \frac{\rho\sigma_X y}{\sigma_Y}\right)^2}{2\sigma_X^2(1-\rho^2)}}\, dx}_{1}$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma_Y} e^{-\frac{y^2}{2\sigma_Y^2}}$$

$$Y \sim N(0, \sigma_Y^2)$$

b) $$\underline{f_{X|Y}(x|y)} = \frac{1}{\sqrt{2\pi}\,\sigma_X\sqrt{1-\rho^2}} e^{-\frac{1}{2\sigma_X^2(1-\rho^2)}\left(x - \frac{\rho\sigma_X y}{\sigma_Y}\right)^2}$$

$$X|Y \sim N\left(\frac{\rho\sigma_X y}{\sigma_Y},\ \sigma_X^2(1-\rho^2)\right)$$