Anushi R. Sadam
Professor Tan Bui Thanh
COE 379L: Introduction to Machine Learning & Data Scienc
29 October 2024

Homework 4

1. Find the mean, covariance of $P(y|x, D)$ w/ Bayesian approach.

$$P(y|x,D) = \int_\theta P(y|x,\theta)\, P(\theta|D)\, d\theta$$

$$P(\theta|D) = \frac{P(D|\theta)\cdot P(\theta)}{P(D)}$$

we assume the prior $P(\theta) = N(0, \sigma^2 I_d)$ (mean 0, variance σ^2), since we have no info about the prior in our test problem. d is the dimension of θ; $\theta \in \mathbb{R}^d$.

$P(D|\theta)$ is the likelihood based on the dataset $D = \{(x^i, y^i)\}_{i=1}^n$ ~ number of samples $= |D|$

$$P(D|\theta) = \prod_{i=1}^{n} P(x^i, y^i|\theta)$$

we know $P(x, y|\theta) = P(y|x,\theta)\cdot P(x|\theta)$

Since $x \perp \theta$ (x is independent of θ), $P(x|\theta) = P(x)$. $P(x)$ is an uniform distribution since we must assume our data collection evenly covers the system.

We assume a gaussian conditional distribution:

$$P(y^i | x^i, \theta) = N(\theta^T x^i, \sigma^2)$$

(x^i: feature vector)

or: (n data samples)

$$P(y | x, \theta) = \prod_{i=1}^{n} N(\theta^T x^i, \sigma^2)$$

(σ^2: variance) assume same noise level for all measurements.

Thus, $P(x, y | \theta) \propto P(y | x, \theta)$

(x: feature matrix, vectors; n: number of data points)

$$\propto \prod_{i=1}^{n} \exp\left(\frac{-1}{2\sigma^2} (y^i - \theta^T x^i)^2\right)$$

$P(x, y | \theta) = P(y | x, \theta) \cdot P(x | \theta)$

mean of $P(y^i | x^i, \theta) = \theta^T x^i$: proven in HW 3.

Since x is independent of θ, $P(x|\theta)$ is $P(x)$.

So, since $P(x)$ is a uniform distribution,

$P(x, y | \theta) \propto P(y | x, \theta)$

normal distribution assumed by our model.

$$P(D|\theta) = P(x, y | \theta) \propto \prod_{i=1}^{n} \exp\left(\frac{-1}{2\sigma^2} (y^i - \theta^T x^i)^2\right)$$

We assume $P(\theta)$ is a gaussian distribution w/ mean $\vec{0}$, variance of each element λ^2:

Now, we assume a prior of mean 0, variance λ^2 I0
(along all dimensions)

in general:

$$P(\vec{\theta}) = \frac{1}{\sqrt{(2\pi)^{|D|} \cdot |\Sigma|^{1/2}}} \exp\left(-\frac{1}{2}(\vec{\theta}-\vec{\mu})^T \Sigma^{-1} (\vec{\theta}-\vec{\mu})\right)$$

where $|D|$ is the number of dimensions for $\vec{\theta}$,

Σ is the covariance matrix,

μ is the mean of $P(\vec{\theta})$.

We know $\mu = \vec{0}$. Now to calculate Σ:

$$\Sigma_{ij} = \text{cov}(\theta_i, \theta_j) \quad \forall i,j \in [1, |D|]$$

$$= \mathbb{E}[(\theta_i - \mathbb{E}[\theta_i])(\theta_j - \mathbb{E}[\theta_j])]$$

$\mathbb{E}[\theta_i] = \mathbb{E}[\theta_j] = 0$ (mean is $\vec{0}$)

$\Rightarrow \Sigma_{ij} = \mathbb{E}[\theta_i \times \theta_j]$

1) $\forall i = j$:

$\Sigma_{ii} = \mathbb{E}[\theta_i^2]$

we know $\text{Var}(\theta_i) = \mathbb{E}[\theta_i^2] - (\underbrace{\mathbb{E}[\theta_i]}_{0})^2 = \lambda^2$

$\Rightarrow \text{Var}(\theta_i) = \mathbb{E}[\theta_i^2] = \lambda^2$

for each θ_i, variance is a constant, λ^2

Thus, $\Sigma_{ii} = \lambda^2$

2) $\forall i \neq j$

$$\Sigma_{ij} = \mathbb{E}[\theta_i \theta_j]$$

we assume independent, random variables for the coefficients θ_i

so, $\Sigma_{ij} = \mathbb{E}[\theta_i] \cdot \mathbb{E}[\theta_j]$

$= \mu_i \cdot \mu_j$

μ_i: mean of θ_i distribution; μ_j: mean of θ_j distribution

since we assume random normal distributions of mean 0 ($\mu_i = 0$, $\mu_j = 0$),

$$\Sigma_{ij} = 0 \cdot 0 = 0.$$

Thus, $\Sigma = \begin{bmatrix} \lambda^2 & & 0 \\ & \lambda^2 & \\ 0 & & \ddots \lambda^2 \end{bmatrix} = \lambda^2 I_{|D|}$

$I_{|D|}$: identity of dimension $|D| \times |D|$

$$\det(\Sigma) = (\lambda^2)^{|D|}$$

$$\Sigma^{-1} = (\lambda^2 I_{|D|})^{-1} = \frac{1}{\lambda^2} I_{|D|}$$

Note $P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)}$

$P(D)$ is simply a normalization value used so that the total probability $\int_D P(\theta|D)\, dD = 1$.

($\Rightarrow P(\theta|D)$ is a bona fide probability distribution)

Then,

$$P(\theta|D) \propto P(D|\theta) \cdot P(\theta)$$

$$= \prod_{i=1}^{n} \exp\left(\frac{-1}{2\delta^2} (y^i - \theta^T x^i)^2\right) \cdot$$

$$\exp\left(\frac{-1}{2\lambda^2} \|\theta\|_2^2\right)$$

$$= \exp\left(-\frac{1}{2}\left(\sum_{i=1}^{n} \left[\frac{1}{\delta^2} (y^i - \theta^T x^i)^2\right] + \frac{1}{\lambda^2} \|\theta\|_2^2\right)\right)$$

$P(\theta|D)$.

we seek to maximize ↑ (feature matrix)

$$P(\theta|D) \propto \exp\left(-\frac{1}{2}\left[(y-x\theta)^2 \cdot \frac{1}{\delta^2} + \frac{1}{\lambda^2}\theta^T\theta\right]\right)$$

$$\propto \exp\left(-\frac{1}{2}\left[\frac{1}{\delta^2}\left(y^Ty - 2y^Tx\theta + \theta^Tx^Tx\theta\right) + \frac{1}{\lambda^2}(\theta^T\theta)\right]\right)$$

$$\propto \exp\left(-\frac{1}{2}\left(\theta^T\underbrace{\left(\frac{x^Tx}{\delta^2} + \frac{I}{\lambda^2}\right)}_{A}\theta - 2\underbrace{\frac{y^Tx}{\delta^2}}_{b^T}\theta + \underbrace{\frac{y^Ty}{\delta^2}}_{C}\right)\right)$$

let $A = \left[\frac{x^Tx}{\delta^2} + \frac{I}{\lambda^2}\right]$ & $b = \frac{x^Ty}{\delta^2}$ & $C = \frac{y^Ty}{\delta^2}$

then,

$$-\frac{1}{2}\left(\theta^TA\theta - 2b^T\theta + C\right) = -\frac{1}{2}\left(\theta^TA\theta - 2b^T\theta + b^TA^{-1}b - b^TA^{-1}b + C\right)$$

$$= -\frac{1}{2}\left((\theta - A^{-1}b)^TA(\theta - A^{-1}b) - b^TA^{-1}b + C\right)$$

F Aside:

$$\theta^TA\theta - 2b^T\theta = \left(\theta^TA\theta - 2b^T\theta + b^TA^{-1}b\right) - b^TA^{-1}b - 2b\theta$$

$$\theta^TA\theta - 2b^T\theta + b^TA^{-1}b = (\theta - A^{-1}b)^TA(\theta - A^{-1}b)$$

since:

LHS expansion:

$$(\theta^T)\ A\theta - (\theta^T)\ A\cdot A^{-1}b - (A^{-1}b)^TA\theta + (A^{-1}b)^TA(A^{-1}b)$$

$(A^{-1}b)^T A = b^T (A^{-1})^T A.$

Note since $I = A^{-1}A = AA^{-1}$; $I^T = (A^{-1}A)^T = (AA^{-1})^T$

$I = I^T = (AA^{-1})^T = (A^{-1})^T A^T$

Note that A is a symmetric matrix $A = mX^TX + nI$ where m, n are constants

$$A^T = ((mX^TX) + nI)^T = (nI)^T + (mX^TX)^T$$

$$= nI^T + m(X^TX)^T = nI + m\,X^T(X^T)^T$$
$$= nI + mX^TX = mX^TX + nI = A.$$

Thus, A is symmetric.

So, $I = (A^{-1})^T A^T = (A^{-1})^T A.$

Then, $(A^{-1}b)^T A = b^T I = b^T.$

Continuing the LHS expansion:

$$\theta^T A\theta - \theta^T b - b^T\theta + (b^T \underbrace{(A^{-1})^T A}_{I} A^{-1} b)$$

$$= \theta^T A\theta - 2b^T\theta + b^T(A^{-1})b \quad 7$$

Thus, $P(\theta|D) \propto \exp\left[-\frac{1}{2}\left((\theta - A^{-1}b)^T A(\theta - A^{-1}b) - b^TA^{-1}b + c\right)\right]$

continuing w/ our exponential factoring:

let $F = (b^TA^{-1}b + c)\frac{1}{2}$

Then in the exponent, we have

$$-\frac{1}{2}(\theta - A^{-1}b)^T A (\theta - A^{-1}b) + F.$$

so:

$$P(\theta|D) \propto \exp\left(-\frac{1}{2}(\theta - A^{-1}b)^T A(\theta - A^{-1}b) + F\right)$$

$$\propto \exp\left(-\frac{1}{2}\left((\theta - A^{-1}b)^T A(\theta - A^{-1}b)\right)\right) \cdot e^F$$

e^F constant w.r.t. θ.

$$\propto \exp\left(-\frac{1}{2}(\theta - A^{-1}b)^T A(\theta - A^{-1}b)\right)$$

multivariate Gaussian expression

w/ mean $\theta^* = A^{-1}b$, $\Sigma_0 = A^{-1}$.

Thus, $P(\theta|D) \propto \exp\left(-\frac{1}{2}(\theta - \theta^*)^T \Sigma_0^{-1} (\theta - \theta^*)\right)$

where $\theta^* = \left[\frac{X^TX}{\delta^2} + \frac{I}{\lambda^2}\right]^{-1} \frac{X^Ty}{\delta^2} = \left[X^TX + I\frac{\delta^2}{\lambda^2}\right]^{-1} X^Ty$

$$\Sigma_0 = \left[\frac{X^TX}{\delta^2} + \frac{I}{\lambda^2}\right]^{-1} = \delta^2\left[X^TX + \frac{\delta^2}{\lambda^2}I\right]^{-1}$$

① Note about scalar inverse pg 7

we recognize θ^* as a least squares solution w/ regularization $\frac{\delta^2}{\lambda^2}$

let $\hat{y} = X\theta$, then we have a least squares problem:

$\theta^* = \arg\min_\theta (y - X\theta)^T(y - X\theta) + \frac{\gamma}{2}\|\theta\|^2$, γ is a regularizer

The solution to the L_2 regularized solution is the same as the loss:

$$J(\theta) = \|y - X\theta\|^2 + \frac{\gamma}{2}\|\theta\|^2.$$

$$\nabla_\theta J(\theta) = X^TX\theta - X^Ty + \gamma\theta = 0.$$

$$\theta(X^TX + \gamma I) = X^Ty$$

positive definite ; $\theta^*_{MAP} = (X^TX + \gamma I)^{-1} X^Ty$

Thus, for $P(\theta|D)$, $\theta^* = \left(X^TX + \frac{\sigma^2}{\lambda^2} I\right)^{-1} X^Ty$

which is simply the MAP solution for θ^* via least squares, w/ regularizer $r = \frac{\sigma^2}{\lambda^2}$.

Note that the mean of $P(\theta|D)$ is θ^*, since

$$P(\theta|D) \propto \exp\left(-\frac{1}{2}(\theta - \theta^*)^T \Sigma_0 (\theta - \theta^*)\right)$$

has the same form as the gaussian w/ mean $\mu = \theta^*$ & covariance Σ_0.

Σ_0, thus, is: $\boxed{\Sigma_0 = \left[\frac{X^TX}{\sigma^2} + \frac{I}{\lambda^2}\right]^{-1} = \sigma^2\left[X^TX + \frac{\sigma^2}{\lambda^2} I\right]^{-1}}$

① $(cA)^{-1} = c^{-1}A^{-1}$ $\forall\ c \in \mathbb{R}$, $A \in \mathbb{R}^{d \times d}$ $d \in \mathbb{R}$

↓ A must be invertible

Proof:

Suppose cA has inverse B:

$(cA)B = I$; $c(AB) = I$; $AB = \frac{1}{c} I$

$A^{-1}AB = \frac{1}{c}A^{-1}I$; $IB = \frac{1}{c}A^{-1}$

$\therefore B = \frac{1}{c}A^{-1} = c^{-1}A^{-1}$

Note that $P(\theta|D) = N(\theta^*_{MAP}, \Sigma_0)$; we can change from $\propto$ to $=$ since we know $P(\theta|D)$ must be a bona fide probability distribution, so, $\int_\theta P(\theta|D)\, d\theta = 1$.

If $P(\theta|D) = c \cdot N(\theta^*_{MAP}, \Sigma_0)$ for some constant (scalar) c, then,

$$\int_\theta P(\theta|D)\, d\theta = \int_{-\infty}^{\infty} c \cdot \frac{\exp\left(\frac{-1}{2}(\theta - \theta^*_{MAP})^T \Sigma_0^{-1} (\theta - \theta^*_{MAP})\right)}{(2\pi)^{d/2} |\Sigma_0|^{1/2}} \times d\theta$$

where d is the dimension of the data

we know, then $\int_\theta P(\theta|D)\, d\theta = c \cdot \int_{-\infty}^{\infty} N(\theta^*_{MAP}, \Sigma_0)\, d\theta$.

per the definition of a normal distribution, we know

$$\int_{-\infty}^{\infty} N(\theta^*_{MAP}, \Sigma_0)\, d\theta = 1$$

(area under normal distribution $= 1$).

Then, $\int_\theta P(\theta|D)\, d\theta = c \cdot 1$. so, $\underline{c = 1}$.

This means $P(\theta|D) = N(\theta^*_{MAP}, \Sigma_0)$.

The coefficient $\frac{1}{(2\pi)^{d/2} |\Sigma_0|^{1/2}}$ comes from the various constants we ignored in our proportion analysis done previously. For example, one of the constants being ignored is $P(D)$ when we solve for $P(\theta|D)$. This will help allow $P(\theta|D)$ to be normalized.

1. cont.

p.3. Now, to find $P(y|x, D)$:

$$P(y|x, D) = \int_\theta P(y|x, \theta)\, P(\theta|D)\, d\theta$$

$$= \int_\theta \prod_{i=1}^{n} N(\theta^T x^i, \delta) \cdot N(\theta_{MAP}, \Sigma_0)\, d\theta$$

note that

$$\prod_{i=1}^{n} N(\theta^T x^i, \delta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\delta^2}} \exp\left(\frac{-1}{2\delta^2}(y^i - \theta^T x^i)^2\right)$$

$$= \frac{1}{(2\pi)^{n/2}\delta^n} \exp\left(\frac{-1}{2\delta^2}(y - X\theta)^T(y - X\theta)\right)$$

where X is the feature matrix.

So, $P(y|x, \theta) \cdot P(\theta|D) =$

$$\frac{1}{(2\pi)^{n/2}\delta^n} \cdot \frac{1}{(2\pi)^{n/2}|\Sigma_0|} \exp\Big(\underbrace{\frac{-1}{2\delta^2}(y - X\theta)^T(y - X\theta) + \frac{-1}{2}(\theta - \theta_{MAP})^T \Sigma_0^{-1} (\theta - \theta_{MAP})}_{M}\Big)$$

let M be the exponent term.

we expand M to get:

$$\frac{-1}{2\delta^2}\left(y^T y - 2y^T X\theta + \theta^T X^T X\theta\right) \frac{-1}{2}\left(\theta^T \Sigma_0^{-1}\theta - 2\theta^T \Sigma_0^{-1}\theta_{MAP} + \theta_{MAP}^T \Sigma_0^{-1}\theta_{MAP}\right)$$

Grouping quadratic terms:

$$M = -\frac{1}{2}\left(\theta^T \left(\frac{x^T x}{\delta^2} + \Sigma_0^{-1} \right) \theta \right.$$

$$\left. -2\theta^T \left(\frac{x^T y}{\delta^2} + \Sigma_0^{-1} \theta_{MAP} \right) \right)$$

$$\underbrace{-\frac{1}{2\delta^2} y^T y - \frac{1}{2} \theta_{MAP}^T \Sigma_0^{-1} \theta_{MAP}}_{\text{constant } C \text{ wrt } \theta}$$

C is a constant so, we can factor out e^C as a constant from the integral of $P(y|x,D)$.

Thus, define $Q = -\frac{1}{2}\left(\theta^T A \theta - 2\theta^T b \right)$

be the exponent in the integral,

where $A = \frac{x^T x}{\delta^2} + \Sigma_0^{-1}$ & $b = \frac{x^T y}{\delta^2} + \Sigma_0^{-1} \theta_{MAP}$.

As before, we can complete the square

to have $Q = -\frac{1}{2}\left((\theta - A^{-1}b)^T A (\theta - A^{-1}b) \right) + \frac{1}{2} b^T A^{-1} b$.

Then, we have: $e^{\downarrow} \propto N(A^{-1}b, A^{-1})$

$$P(y|x,D) = \int_\theta \frac{1}{(2\pi)^{n/2} \delta^n |\Sigma_0|} \exp(Q)\, d\theta \cdot e^C$$

now with some factoring we see:

$$P(y|x,D) = \gamma \int_\theta \underbrace{N(A^{-1}b, A^{-1})}_{\text{function of } \theta} \cdot \underbrace{e^{\frac{1}{2} b^T A^{-1} b}}_{\text{not function of } \theta, \text{constant}} d\theta$$

where $\gamma = \dfrac{e^c}{(2\pi)^{\frac{n}{2}} \delta^n |\Sigma_0|}$.

then,

$$P(y|x,D) = \gamma e^{\frac{1}{2} b^T A^{-1} b} \int_\theta \underbrace{N(A^{-1}b, A^{-1})\, d\theta}_{\text{gaussian distribution integral} = 1.}$$

$$= \gamma e^{\frac{1}{2} b^T A^{-1} b}$$

No where to proceed γ should lead to our Gaussian function for the posterior predictive distribution

1. part 3 cont.

We see pure integration does not lead to reasonable results - we are stuck.

So, we instead use normal distribution properties to compute the posterior predictive distribution:

Our Bayesian regression model:

$$y = X\theta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I_n)$$

$y \in \mathbb{R}^n$, n: number of data points.

$X \in \mathbb{R}^{n \times d}$

$\theta \in \mathbb{R}^d \implies P(y|x,\theta) = N(\theta^T x, \sigma^2)$, where $\theta^T x = x^T\theta$

Our posterior:

$$P(\theta|D) \sim N(\theta_{MAP}, \Sigma_0)$$

Since $P(y|x,D) = \int_\theta P(y|x,\theta) \cdot P(\theta|D)\, d\theta$

or our posterior predictive model is the convolution of 2 gaussians.

So, $P(y|x, D)$ is a gaussian distribution.

Recall the linear transformation rule on a normal distribution:

$$y = Ax, \quad x \sim N(\mu, \Sigma), \quad P(y) = N(y \mid A\mu, A\Sigma A^T)$$

(normal w.r.t. y)

Proof:

$X \sim N(\mu, \Sigma)$, $\mathbb{E}[X] = \mu$.

$Y = AX$. $\mathbb{E}[Y] = \mathbb{E}[AX] = A\,\mathbb{E}[X] = A\mu$.

So, mean of Y, $\mathbb{E}[Y] = A\mu$.

$$Cov(X) = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T] = \Sigma$$

$$Cov(Y) = Cov(AX) = \mathbb{E}[(AX - \mathbb{E}[AX])(AX - \mathbb{E}[AX])^T]$$

$$= \mathbb{E}[(AX - A\mu)(Ax - A\mu)^T]$$

$$= \mathbb{E}[A(x - \mu) \cdot (A(x-\mu))^T]$$

$$= \mathbb{E}[A(x-\mu)(x-\mu)^T A^T]$$

$$= A\,\mathbb{E}[(x-\mu)(x-\mu)^T]A^T$$

$$= A \cdot \Sigma A^T.$$

Now, going back to $P(y^*|x^*, D)$: where x^*, y^* is the new point(s) being predicted:

$$\mathbb{E}[y^*|x^*, D] = \mathbb{E}[\underbrace{\mathbb{E}[y^*|x^*, \theta]}_{\text{from } P(y|x,\theta)}|D]$$

$$= \mathbb{E}[(x^*)^T \theta | D]$$

$\Downarrow$

per our model, $P(y|x, \theta)$

$$= (x^*)^T\,\mathbb{E}[\theta|D] = x^{*T}\theta_{MAP}.$$

per prior, $\Downarrow = \theta_{MAP}$

Thus, the mean of $P(y^* | x^*, D) = (x^*)^T \theta_{MAP}$.

Now, for the posterior predictive variance:

$$\text{Var}(y^* | x^*, D) = \mathbb{E}[\text{var}(y^* | x^*, \theta) | D] + \text{var}(\mathbb{E}[y^* | x^*, \theta] | D)$$

} Total variance Law

$\mathbb{E}[\text{var}(y^* | x^*, \theta) | D] = \mathbb{E}[\delta^2 | D] = \delta^2$

δ^2, per $P(y|x,\theta)$ model

$\text{Var}(\mathbb{E}[y^* | x^*, \theta] | D) = \text{Var}((x^*)^T \theta_{MAP} | D)$

mean of $P(y^* | x^*, D)$

$= x^{*T} \text{cov}(\theta | D)\, x^*$ ← linear transformation rules as previously explained.

$= x^{*T} \Sigma_0 x^*$

Thus, $\text{Var}(y^* | x^*, D) = \delta^2 + x^{*T} \Sigma_0 x^*$

So,

$$P(y^* | x^*, D) = N(x^{*T} \theta_{MAP}, x^{*T} \Sigma_0 x^* + \delta^2)$$

or:

$$P(y | x, D) = N(x^T \theta_{MAP}, x^T \Sigma_0 x + \delta^2)$$

2. Prove: $\mu_K = \frac{\sum_{i:y^i=K} x^i}{n_K}$, $\Sigma_K = \frac{\sum_{i:y^i=K} (x^i-\mu_K)(x^i-\mu_K)^T}{n_K}$

for $x^i \in \mathbb{R}^n$, $\mu \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{n \times n}$, n is number of features (dimension) in training sample x^i.

Let there be K classes. we maximize:

$$\max_{\mu_K, \Sigma_K} \sum_{i, y^i=K} \log P(x^i | y^i, \mu_K, \Sigma_K) =$$

$$= \max_{\mu_K, \Sigma_K} \sum_{i:y^i=K} \log (N(x^i | \mu_K, \Sigma_K)) = \max_{\mu_K, \Sigma_K} L(\mu_K, \Sigma_K)$$

let n_K be the number of data samples (x^i, y^i) such that $y^i = K$ (class K, $K \in [1, K]$)

$$N(x^i | \mu_K, \Sigma_K) = \frac{1}{(2\pi)^{n/2} |\Sigma_K|^{1/2}} \exp\left(-\frac{1}{2}(x^i - \mu_K)^T \Sigma_K^{-1} (x^i - \mu_K)\right)$$

$$\log N(x^i | \mu_K, \Sigma_K) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_K| - \frac{1}{2}(x^i - \mu_K)^T \Sigma_K^{-1} (x^i - \mu_K)$$

Thus, our problem likelihood $L(\mu_K, \Sigma_K)$ is

$$L(\mu_K, \Sigma_K) = -\frac{n_K n}{2}\log(2\pi) - \frac{n_K}{2}\log|\Sigma_K| - \sum_{i:y^i=K} \frac{1}{2}(x^i - \mu_K)^T \Sigma_K^{-1} (x^i - \mu_K)$$

To maximize $L(\mu_K, \Sigma_K)$, we set $\frac{\partial L}{\partial \mu_K}, \frac{\partial L}{\partial \Sigma_K} = 0$.

1) $\frac{\partial L}{\partial \mu_K} = \frac{-1}{2} \sum_{i:y^i=K} \frac{\partial}{\partial \mu_K} (x^i - \mu_K)^T \Sigma_K^{-1} (x^i - \mu_K)$

we know $\frac{d}{dx} x^T A x = 2Ax$.

(shown in HW1).

So, with the chain rule,

$$\frac{\partial L}{\partial \mu_k} = \frac{1}{2} \sum_{i:y^i=k} \Sigma_k^{-1} (x^i - \mu_k)(-2)$$

$$= \sum_{i:y^i=k} \Sigma_k^{-1} (x^i - \mu_k)$$

$$= \Sigma_k^{-1} \cdot \sum_{i:y^i=k} (x^i - \mu_k)$$

$$= \Sigma_k^{-1} \left(-n_k \mu_k + \sum_{i:y^i=k} x^i \right)$$

Setting $\frac{\partial L}{\partial \mu_k} = 0$, we have:

$$\Sigma_k^{-1} \left(\sum_{i:y^i=k} x^i - n_k \mu_k \right) = \vec{0}$$

Since Σ_k^{-1} is invertible $(\Sigma_k^{-1})^{-1} = \Sigma_k$ exists & Σ_k^{-1} is not a null matrix ($\Sigma_k \Sigma_k^{-1} = I$, not null),

we have $\underbrace{\Sigma_k \Sigma_k^{-1}}_{I} \left(\sum_{i:y^i=k} x^i - n_k \mu_k \right) = \Sigma_k \cdot \vec{0}$

$\exists$

Thus, $\sum_{i:y^i=k} x^i - n_k \mu_k = \vec{0}$.

So, $$\boxed{\mu_k = \frac{\sum_{i:y^i=k} x^i}{n_k}}$$

Now, setting $\frac{\partial L}{\partial \Sigma_k} = 0$:

$\partial L / \partial \Sigma_k =$

$$\frac{\partial}{\partial \Sigma_k} \left(\frac{n_k}{2} \log |\Sigma_k| - \frac{1}{2} \sum_{i:y^i=k} (x^i - \mu_k)^T \Sigma_k^{-1} (x^i - \mu_k) \right)$$

$$\frac{\partial}{\partial \Sigma_k} \frac{n_k}{2} \log |\Sigma_k| = \frac{n_k}{2} \frac{\partial}{\partial \Sigma_k} \log |\Sigma_k|$$

$$\frac{\partial}{\partial \Sigma_k} \log |\Sigma_k| = \frac{1}{|\Sigma_k|} \cdot \frac{\partial}{\partial \Sigma_k} |\Sigma_k|$$

Math Derivations:

we know Σ_K is a square, invertible matrix.

we know $C_{ij} = (-1)^{i+j} M_{ij}$; $M_{ij} = \det(\Sigma_{ij})$

where C is the cofactor at row i, column j & M is the minor.

Then, $|\Sigma_K| = \sum_{j=1}^{n} \Sigma_{ij} C_{ij}$

$= \sum_{j=1}^{n} \Sigma_{lj} C_{lj}$ ✓ for any $\forall l \in [1, d]$

let $l = a$

$$\frac{\partial |\Sigma|}{\partial \Sigma_{ab}} = \frac{\partial}{\partial \Sigma_{ab}} \left(\sum_{j=1}^{n} \Sigma_{aj} C_{aj} \right)$$

$$= \sum_{j=1}^{n} \frac{\partial}{\partial \Sigma_{ab}} \Sigma_{aj} C_{aj}$$

$= C_{ab} = C_{ba} \rightarrow i+j = a+b$ for C_{ab} & C_{ba}.

$M_{ij} = M_{ji}$ since Σ is symmetric.

$|\Sigma| = \Sigma \cdot C$ $\quad \Sigma \cdot \Sigma^{-1} = I_d$

$\Sigma = \frac{|\Sigma|}{C}$ $\quad \Sigma^{-1} = \frac{I_d}{\Sigma} = \frac{C}{|\Sigma|}$; $C = |\Sigma| \Sigma^{-1}$

So, $\frac{\partial}{\partial \Sigma} |\Sigma| = C = |\Sigma| \Sigma^{-1}$.

Or, $\frac{\partial}{\partial \Sigma_K} |\Sigma_K| = |\Sigma_K| \Sigma_K^{-1}$

so, $\frac{\partial}{\partial \Sigma_k} \log|\Sigma_k| = \frac{1}{|\Sigma_k|} |\Sigma_k| \Sigma_k^{-1} = \Sigma_k^{-1}$

F Aside: $\frac{\partial}{\partial A} A^{-1} = ?$

$A^{-1}A = I \Rightarrow \frac{\partial}{\partial A}(A^{-1}A) = 0$.

W/ product rule:

$\frac{\partial A^{-1}}{\partial A} \cdot A + \frac{\partial A}{\partial A} \cdot A^{-1} = 0$

$\frac{\partial A^{-1}}{\partial A} \cdot A + A^{-1} = 0 \Rightarrow \frac{\partial A^{-1}}{\partial A} A = -A^{-1}$

$\frac{\partial A^{-1}}{\partial A} A A^{-1} = -A^{-1}A^{-1}$; $\frac{\partial A^{-1}}{\partial A} = -A^{-1}A^{-1}$

$\frac{\partial}{\partial A} v^T A^{-1} v = ?$ let $f(A) = v^T A^{-1} v$

$\frac{\partial}{\partial A}$ ↑ vector

$\frac{\partial}{\partial A} f(A) = v^T \frac{\partial}{\partial A} A^{-1} v = v^T(-A^{-1} \cdot A^{-1}) v$

$\frac{\partial}{\partial A}(v^T A^{-1} v) = -v^T A^{-1} A^{-1} v$

$\frac{\partial L}{\partial \Sigma_k} = -\frac{n_k}{2} \frac{\partial}{\partial \Sigma_k}(\log|\Sigma_k|) - \frac{1}{2} \sum_{i: y^i = k} \frac{\partial}{\partial \Sigma_k} (x^i - \mu_k)^T \Sigma_k^{-1} (x^i - \mu_k)$

$\frac{\partial L}{\partial \Sigma_k} = -\frac{n_k}{2} \Sigma_k^{-1} + \frac{1}{2} \sum_{i: y^i = k} (x^i - \mu_k)^T \Sigma_k^{-1} \Sigma_k^{-1} (x^i - \mu_k)$

† Aside: $v^T A^{-1} A^{-1} v = A^{-1} v v^T A^{-1}$ when A is symmetric?

↳ $v \in \mathbb{R}^d$, $A \in \mathbb{R}^{d \times d}$

let $p = A^{-1}v$. $p^T = v^T (A^{-1})^T$

for $A = A^T$:

$A^{-1}A = I$;

$I = (A^T)^{-1}(A^T) = (A^T)^{-1} A = I$; $(A^T)^{-1} = A^{-1}$

So, $p^T = v^T A^{-1}$

Thus, $v^T A^{-1} A^{-1} v = p^T p = \sum_{i=1}^{d} (p_i)^2 = \| A^{-1} v \|^2$

Now, we use proof by contradiction:

we evaluate the RHS of the hypothesis:

$A^{-1} v \cdot v^T A^{-1} = p \cdot p^T = \sum_{i=1}^{d} (p_i)^2 = \| A^{-1} v \|^2$

Thus for the hypothesis to be false we must have $\| A^{-1} v \|^2 \neq \| A^{-1} v \|^2$ which is obviously false.

Both sides can be simplified to the same value so

$v^T A^{-1} A^{-1} v = A^{-1} v v^T A^{-1}$ 7

Thus, we have, setting $\frac{\partial L}{\partial \Sigma_k} = 0$,:

$\frac{n_k}{2} \Sigma_k^{-1} = \frac{1}{2} \sum_{i : y^i = k} (x^i - \mu_k)^T \Sigma_k^{-1} \Sigma_k (x^i - \mu_k)$

Scalar ↳ $\frac{n_k}{2} \Sigma_k^{-1} = \frac{1}{2} \sum_{i : y^i = k} \Sigma_k^{-1} (x^i - \mu_k)(x^i - \mu_k)^T \Sigma_k$

$$\Sigma_k^{-1} \frac{n_k}{2} = \frac{1}{2} \sum_{i:y^i=k} \Sigma_k^{-1} (x^i - \mu_k)(x^i - \mu_k)^T \Sigma_k^{-1}$$

multiplying both sides by Σ_k from the left:

$$\Sigma_k \Sigma_k^{-1} \frac{n_k}{2} = \frac{1}{2} \sum_{i:y^i=k} \Sigma_k \Sigma_k^{-1} (x^i - \mu_k)(x^i - \mu_k)^T \Sigma_k^{-1}$$

$$n_k = \sum_{i:y^i=k} (x^i - \mu_k)(x^i - \mu_k)^T \Sigma_k^{-1}$$

Note that since Σ_k is symmetric we can multiply Σ_k from the right on equations. In the above equation we have a scalar on the left side so, this property isn't necessarily required.

$$n_k \Sigma_k = \sum_{i:y^i=k} (x^i - \mu_k)(x^i - \mu_k)^T \Sigma_k^{-1} \Sigma_k$$

$$\Rightarrow$$

$$\boxed{\Sigma_k = \frac{\sum_{i:y^i=k} (x^i - \mu_k)(x^i - \mu_k)^T}{n_k}}$$

1) Derive $\psi_{jK} = \frac{n_{jK}}{n_K}$

ψ_{jK}: $P(x=1 \mid y=K)$, probability of seeing word that belongs to vocabulary set $\phi(x)$, given restriction to category K (i.e spam email category).

n_{jK}: frequency of word j in class K

n_K: number of documents in class k.

for dataset $\mathcal{D} = \{(x^i, y^i) \mid i = 1, 2 \cdots n\}$ we optimize/maximize the log-likelihood $\log(P_\theta(x,y))$ w/ maximum likelihood:

$$\theta = \{\phi_1, \phi_2 \cdots \phi_K, \psi_{11}, \psi_{21} \cdots \psi_{dK}\}$$

$$\theta^* = \arg\max_\theta \sum_{i=1}^{n} \log P_\theta(x^{(i)}, y^{(i)}) = \arg\max_\theta \ell(\theta)$$

$$P_\theta(x,y) = P_\theta(x \mid y) \cdot P_\theta(y).$$

Note:

$P_\theta(x_j = 1 \mid y = K) = \text{Bernoulli}(\psi_{jK})$

$P_\theta(y) = \text{Categorical}(\phi_1 \cdots \phi_K)$

so,

$$\ell(\theta) = \underbrace{\sum_{i=1}^{n} \log P_\theta(x^{(i)}, y^{(i)})}_{\text{only dependent on } \vec{\psi}} + \underbrace{\sum_{i=1}^{n} \log P_\theta(y^{(i)})}_{\text{only dependent on } \vec{\phi}}$$

$$= \sum_{k=1}^{K} \sum_{j=1}^{d} \sum_{i:y^i=k} \log P(x^{(i)} \mid \psi_{jK}) + \sum_{i=1}^{n} \log P(y^{(i)} \mid \vec{\phi})$$

We know:

$$P(x_j^{(i)} \mid \psi_{jk}) = \psi_{jk}^{x_j^{(i)}} (1-\psi_{jk})^{1-x_j^{(i)}}$$

so,

$$l(\theta) = \sum_{k=1}^{K} \sum_{j=1}^{d} \sum_{i:y^i=k} \left(x_j^{(i)} \log \psi_{jk} + (1-x_j^{(i)})(1-\psi_{jk}) \right) + \sum_{i=1}^{n} \log P(y^i \mid \vec{\phi})$$

to find the maxima (local) of $l(\theta)$, we set $\frac{\partial l}{\partial \psi_{jk}} = 0$ & to get $\Leftrightarrow$

$$\frac{\partial l}{\partial \phi} = 0.$$

1) $$\frac{\partial l}{\partial \psi_{jk}} = \sum_{i:y^{(i)}=k} \frac{\partial}{\partial \psi_{jk}} \left[x_j^{(i)} \log \psi_{jk} + (1-x_j^{(i)})(1-\psi_{jk}) \right]$$

$=0$

$$\frac{\partial l}{\partial \psi_{jk}} = \sum_{i:y^i=k} \frac{1}{\psi_{jk}} x_j^{(i)} - \frac{1}{1-\psi_{jk}} (1-x_j^{(i)})$$

$$\Rightarrow \sum_{i:y^i=k} \frac{x_j^{(i)}}{\psi_{jk}} - \frac{1-x_j^{(i)}}{1-\psi_{jk}} = 0.$$

$$\sum_{i:y^i=k} \frac{x_j^{(i)}}{\psi_{jk}} = \sum_{i:y^i=k} \frac{1-x_j^{(i)}}{1-\psi_{jk}}$$

using our definition for n_{jk} and rearranging:

$$\sum_{i:y^{(i)}=k} x_j^{(i)} (1-\psi_{jk}) = \sum_{i:y^{(i)}=k} \left(1-x_j^{(i)}\right) \psi_{jk}$$

recall that n_{jk} is the number of times $x_j^{(i)}=1$ (vs. $x_j^{(i)}=0$).

$n_k = \sum_j n_{jk}$, total # of observations of class k.

So, we have:

$$n_{jk}(1-\psi_{jk}) = (n_k - n_{jk})\psi_{jk}$$

$$\Rightarrow n_{jk} = n_k \psi_{jk} \quad \text{or} \quad \boxed{\psi_{jk} = \frac{n_{jk}}{n_k}}$$