CS 189 Introduction to Machine Learning
Fall 2017 **HW4**

This homework is due **Friday, September 22 at 10pm.**

1 Getting Started

You may typeset your homework in latex or submit neatly handwritten and scanned solutions. Please make sure to start each question on a new page, as grading (with Gradescope) is much easier that way! Deliverables:

1. Submit a PDF of your writeup to assignment on Gradescope, "HW[n] Write-Up"
2. Submit your test set evaluation results, "HW[n] Test Set".

After you've submitted your homework, be sure to watch out for the self-grade form.

(a) Before you start your homework, write down your team. Who else did you work with on this homework? List names and email addresses. In case of course events, just describe the group. How did you work on this homework? Any comments about the homework?

None
Comments: lecture 6 notes should not be released too late

(b) Please copy the following statement and sign next to it:

I certify that all solutions are entirely in my words and that I have not looked at another student's solutions. I have credited all external sources in this write up.

I certify that all solutions are entirely in my words & that I have not looked at another stud's solutions. I have credited all externals sources in this write up

Problem # 2

a) X has multivariate Gaussian distribution with mean $\mu \in R^d$ and covariance matrix $\Sigma \in R^{d\times d}$

$$p(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} \|\Sigma\|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

Drawing n i.i.d sample $x_1, x_2, \dots, x_n$

$$\prod_1^n p(x_i \mid \mu, \Sigma) = \prod_1^n \left(\frac{1}{(2\pi)^{d/2}\|\Sigma\|^{1/2}}\right)^n e^{-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)}$$

log likelihood:

$$-n\log\left[(2\pi)^{d/2}\|\Sigma\|^{1/2}\right] - \frac{1}{2}\sum_1^n (x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$

b.) Take derivative wrt μ:

$$\frac{\partial(\,)}{\partial\mu} = -0 - \frac{\partial}{\partial\mu}\left[\frac{1}{2}\sum_1^n(\mu-x_i)^T\Sigma^{-1}(\mu-x_i)\right]$$

$$= -\sum_1^n (\mu - x_i)^T\Sigma^{-1}$$

$$= \left(n\mu - \sum_1^n x_i\right)^T \Sigma^{-1} = 0$$

$$\Rightarrow \mu = \frac{1}{n}\sum_1^n x_i$$

$\frac{\partial x^TAx}{\partial x} = x^T(A+A^T)$

$= 2x^TA$ if A sym

$\frac{\partial u^Tv}{\partial x} = u^T\frac{\partial v}{\partial x} + v^T\frac{\partial u}{\partial x}$

$\frac{\partial Ax}{\partial x} = A \quad \frac{\partial x^TA}{\partial x} = A^T$

Take derivative w.r.t Σ^{-1}

$$\frac{\partial}{\partial\Sigma^{-1}}\left[\frac{n}{2}\log\left((2\pi)^{-d/2}\|\Sigma\|^{-1}\right) - \frac{1}{2}\sum_1^n(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right]$$

$\frac{\partial \ln\|X\|}{\partial X} = X^{-1}$

$$= \frac{n}{2}(\Sigma^{-1})^{-1} - \frac{1}{2}\sum_1^n(x_i-\mu)^T(x_i-\mu) = 0$$

$$\Rightarrow \Sigma = \frac{1}{n}\sum_1^n (x_i-\mu)^T(x_i-\mu)$$

$\left(\frac{\partial a^TXb}{\partial X} = ab^T \quad \frac{\partial a^TX^Tb}{\partial X} = ba^T\right)$

(c) See appendix

Problem # 3

(a) $z \sim N(0,1) \Rightarrow P(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$

$Y = w^T X + z \Rightarrow Y \sim N(w^T X, 1)$ given X & w

$$\Rightarrow P(Y|X,w) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(Y-w^TX)^2}{2}}$$

(b)

$$P(w) = \frac{1}{(2\pi)^{d/2} \|\Sigma\|^{1/2}} e^{-\frac{1}{2} w^T \Sigma^{-1} w}$$

$$P(w|Y,X) = \frac{P(Y|w,X)\, \cancel{P(w|X)}}{\int_Y P(Y|w,X)\, \underbrace{\cancel{P(w|X)}}_{P(w)}}$$

$$= \text{const} \times e^{-\frac{(Y-w^TX)^2}{2} - \frac{1}{2} w^T \Sigma^{-1} w}$$

$$= \text{const} \times e^{-\frac{Y^2 - 2Yw^TX + w^TXX^Tw}{2} - \frac{1}{2} w^T \Sigma^{-1} w}$$

$$= \text{const} \times e^{-\frac{1}{2} w^T (\Sigma^{-1} + XX^T) w + Yw^TX}$$

Given n training data points $(X_1, Y_1), \dots, (X_n, Y_n)$

$$P(w | Y_1, Y_2, \dots, Y_n ; X_1, X_2, \dots, X_n)$$

$$= P(w|Y_1, X_1)\, P(w|Y_2, X_2) \dots P(w|Y_n, X_n)$$

$$= \text{const} \times e^{-\frac{1}{2} w^T (n\Sigma^{-1} + \sum_1^n X_i X_i^T) w + \sum_1^n Y_i X_i^T w}$$

The mean of w is : $(n\Sigma^{-1} + \sum_1^n X_i X_i^T)^{-1} \sum_1^n X_i Y_i$

(c) BONUS : skip

(d) see appendix

Observation : when we have more training data point, the area of contours shrink

→ prediction more precise

(e) see appendix

Problem #4

(a) $\operatorname{rank}(A + \hat{A}) = d$

We know that we have to find $\hat{A}$ and $\hat{\vec{y}}$ such that:

$$[A + \hat{A}, \vec{y} + \hat{\vec{y}}] \begin{bmatrix} \vec{w} \\ -1 \end{bmatrix} = 0$$

$$(A + \hat{A})\vec{w} - (\vec{y} + \hat{\vec{y}}) = 0$$

$$\Rightarrow \quad \vec{y} + \hat{\vec{y}} = (A + \hat{A})\vec{w}$$

The column $\vec{y} + \hat{\vec{y}}$ is linearly dependent on the columns of $A + \hat{A}$ $\Rightarrow$ adding column $\vec{y} + \hat{\vec{y}}$ to $(A + \hat{A})$ does not change the rank of $A + \hat{A}$, i.e

$$\operatorname{rank}([A + \hat{A}, \vec{y} + \hat{\vec{y}}]) = \operatorname{rank}(A + \hat{A}) = d$$

(b)

$$[A, \vec{y}] = U \Sigma V^T = \begin{bmatrix} U_{xx} & U_{xy} \\ U_{yx} & U_{yy} \end{bmatrix} \begin{bmatrix} \Sigma_{1,\dots,d} & \\ & \sigma \end{bmatrix} \begin{bmatrix} V_{xx} & V_{xy} \\ V_{yx} & V_{yy} \end{bmatrix}^T$$

$$[A + \hat{A}, \vec{y} + \hat{\vec{y}}] = \begin{bmatrix} U_{xx} & U_{xy} \\ U_{yx} & U_{yy} \end{bmatrix} \begin{bmatrix} \Sigma_{1,\dots,d} & \\ & 0 \end{bmatrix} \begin{bmatrix} V_{xx} & V_{xy} \\ V_{yx} & V_{yy} \end{bmatrix}^T$$

$$\Rightarrow [\hat{A}, \hat{\vec{y}}] = \begin{bmatrix} U_{xx} & U_{xy} \\ U_{yx} & U_{yy} \end{bmatrix} \begin{bmatrix} 0 & \\ & -\sigma \end{bmatrix} \begin{bmatrix} V_{xx} & V_{xy} \\ V_{yx} & V_{yy} \end{bmatrix}^T$$

We try to write $[\hat{A}, \hat{\vec{y}}]$ in term of $[A, \vec{y}]$

$$[\hat{A}, \hat{\vec{y}}] = [A, \vec{y}] \times ?$$

$$? = [A, \vec{y}]^{-1} [\hat{A}, \hat{\vec{y}}]$$

$$= V^{-T} \Sigma^{-1} \cancelto{I}{U^{-1} U} \begin{bmatrix} 0 & \\ & -\sigma \end{bmatrix} V^T$$

$$= \cancelto{V}{V^{-T}} \begin{bmatrix} \Sigma^{-1}_{1,\dots,d} & \\ & \sigma^{-1} \end{bmatrix} \begin{bmatrix} 0 & \\ & -\sigma \end{bmatrix} V^T$$

$$= \begin{bmatrix} V_{xx} & V_{xy} \\ V_{yx} & V_{yy} \end{bmatrix} \underbrace{\begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} V_{xx}^T & V_{yx}^T \\ V_{xy}^T & V_{yy}^T \end{bmatrix}}_{\begin{bmatrix} 0 & 0 \\ -V_{xy}^T & -V_{yy}^T \end{bmatrix}} = -\begin{bmatrix} V_{xy} V_{xy}^T & V_{xy} V_{yy}^T \\ V_{yy} V_{xy}^T & V_{yy} V_{yy}^T \end{bmatrix}$$

We have

$$[\hat{A}, \hat{\vec{y}}] = -[A, \vec{y}] \begin{bmatrix} V_{xy} V_{xy}^T & V_{xy} V_{yy}^T \\ V_{yy} V_{xy}^T & V_{yy} V_{yy}^T \end{bmatrix}$$

We multiply both sides with $\begin{bmatrix} V_{xy} \\ V_{yy} \end{bmatrix}$

$$[\hat{A}, \hat{\vec{y}}] \begin{bmatrix} V_{xy} \\ V_{yy} \end{bmatrix} = -[A, \vec{y}] \begin{bmatrix} V_{xy} V_{xy}^T V_{xy} + V_{xy} V_{yy}^T V_{yy} \\ V_{yy} V_{xy}^T V_{xy} + V_{yy} V_{yy}^T V_{yy} \end{bmatrix}$$

we know that $V^T V = I \Rightarrow V_{xy}^T V_{xy} + V_{yy}^T V_{yy} = I$

$$\Rightarrow [\hat{A}, \hat{\vec{y}}] \begin{bmatrix} V_{xy} \\ V_{yy} \end{bmatrix} = -[A, \vec{y}] \begin{bmatrix} V_{xy} \\ V_{yy} \end{bmatrix}$$

$$\Rightarrow [A + \hat{A}, \hat{\vec{y}} + \vec{y}] \underbrace{\begin{bmatrix} V_{xy} \\ V_{yy} \end{bmatrix}}_{\vec{x}} = 0$$

solution $\vec{x} = \begin{bmatrix} V_{xy} \\ V_{yy} \end{bmatrix}$

Multiply $\vec{x}$ with $(-V_{yy}^{-1})$ to normalize $\vec{x}$ in the form $\begin{bmatrix} \vec{w} \\ -1 \end{bmatrix}$

$$\begin{bmatrix} -V_{xy} V_{yy}^{-1} \\ -1 \end{bmatrix} = \begin{bmatrix} \vec{w} \\ -1 \end{bmatrix}$$

$$\Rightarrow \vec{w} = -V_{xy} V_{yy}^{-1}$$

(c) See appendix

(d) See appendix

(e) Scale $= 1/384$

Explanation

$$\vec{f}(\vec{n}) = \sum_{i=1}^{9} \vec{r}_i \, L_i(\vec{n})$$

$\vec{f}(\vec{n}) \times$ scale leads to $L_i(\vec{n})$ / scale and vice versa to keep $\vec{r}_i$ constant.

Problem # 5

$$Y = X_1 + X_2 + Z$$

$$X_1, X_2 \sim N(0,5) \qquad Z \sim N(0,1)$$

prove the theoretical error is

$$5(\hat{w}_1 - 1)^2 + 5(\hat{w}_2 - 1)^2 + 1$$

Solution :

True data $\quad Y = X_1 + X_2 + Z$

model $\quad \hat{Y} = \hat{w}_1 X_1 + \hat{w}_2 X_2$

Error : $\quad \hat{Y} - Y$

$$= X_1(\hat{w}_1 - 1) + X_2(\hat{w}_2 - 1) - Z$$

The expected error

$$E[(\hat{Y} - Y)^2]$$

$$= E[X_1^2(\hat{w}_1 - 1)^2] + E[X_2^2(\hat{w}_2 - 1)^2] + E(Z^2)$$

$$+ \cancelto{0}{E[X_1 X_2 (\hat{w}_1 - 1)(\hat{w}_2 - 1)]} \, *$$

$$+ \cancelto{0}{E[X_1(\hat{w}_1 - 1) Z]} + \cancelto{0}{E[X_2(\hat{w}_2 - 1) Z]}$$

$$= \cancelto{\sigma_{x_1}^2}{E[X_1^2]}(\hat{w}_1 - 1)^2 + \cancelto{\sigma_2^2}{E[X_2^2]}(\hat{w}_2 - 1)^2 + \cancelto{\sigma_z^2}{E(Z^2)}$$

$$= 5(\hat{w}_1 - 1)^2 + 5(\hat{w}_2 - 1)^2 + 1$$