6.867 Machine Learning

Solutions for Problem Set 1

Monday, September 22

Part 1: Least-Squares Regression

Problem 1

(1-1) (5pts) The sample covariance between the (least-squares) prediction error $\hat{e} = y - \hat{w}'\phi(x)$ and the k-th feature $\phi_k(x)$ is

$$\tilde{\sigma}(\hat{e}, \phi_k) = \frac{1}{n}\sum_i (\hat{e}_i - \bar{e})(\phi_k(x_i) - \bar{\phi}_k) \tag{1}$$

where $\hat{e}_i = y_i - \hat{w}'\phi(x_i)$; $\bar{e} = \frac{1}{n}\sum_i \hat{e}_i$ and $\bar{\phi}_k = \frac{1}{n}\sum_i \phi_k(x_i)$. Minimizing $\sum_i e_i^2$ w.r.t. w means that the least-squares predictions satisfy $\sum_i \hat{e}_i \phi_k(x_i) = 0$ for all k. Setting $\phi_1(x) = 1$ implies $\sum_i \hat{e}_i = 0$ so that $\bar{e} = 0$. Then,

$$\begin{aligned}
\tilde{\sigma}(\hat{e}, \phi_k) &= \frac{1}{n}\sum_i \hat{e}_i(\phi_k(x_i) - \bar{\phi}_k) && (2)\\
&= \frac{1}{n}\left\{\left(\sum_i \hat{e}_i\phi_k(x_i)\right) - \left(\sum_i \hat{e}_i\right)\bar{\phi}_k\right\} && (3)\\
&= \frac{1}{n}\left\{0 - 0 \times \bar{\phi}_k\right\} && (4)\\
&= 0 && (5)
\end{aligned}$$

Hence, the optimal linear least-squares predictor based upon features $\phi_1(x) = 1, \phi_2(x), \ldots, \phi_d(x)$ generates prediction errors which are uncorrelated with each of those features.

(1-2) (5pts) Let $\psi(x) = w'\phi(x)$. First, note that ψ is "orthogonal" to the least-squares prediction error $\hat{e} = y - \hat{w}'\phi(x)$ in the sense that

$$\begin{aligned}
\sum_i \hat{e}_i\psi(x_i) &= \sum_{i=1}^{n} \hat{e}_i \left(\sum_{k=1}^{d} w_k\phi_k(x_i)\right) && (6)\\
&= \sum_k w_k \left(\sum_i \hat{e}_i\phi_k(x_i)\right) && (7)\\
&= \sum_k w_k \times 0 && (8)\\
&= 0 && (9)
\end{aligned}$$

If $\phi_1(x) = 1$, then orthogonality implies ψ is uncorrelated with the prediction error as shown below.

$$\begin{aligned}
\tilde{\sigma}(\hat{e}, \psi) &= \frac{1}{n} \sum_i \hat{e}_i (\psi(x_i) - \bar{\psi}) && (10) \\
&= \frac{1}{n} \left\{ \left(\sum_i \hat{e}_i \psi(x_i) \right) - \left(\sum_i \hat{e}_i \right) \bar{\psi} \right\} && (11) \\
&= \frac{1}{n} \left\{ 0 - 0 \times \bar{\psi} \right\} && (12) \\
&= 0 && (13)
\end{aligned}$$

(1-3) (5pts) Given the original data $\{(x_i, y_i), i = 1, ..., n\}$ and specified features $\phi(x)$, we compute the least-squares parameters $\hat{\mathbf{w}} = (X'X)^{-1}X'\mathbf{y}$ and associated prediction errors $\hat{e}_i = y_i - \hat{\mathbf{w}}'\phi(x_i)$. Now, consider the new "data" $\{(x_i, \tilde{y}_i = e_i), i = 1, ..., n\}$. Let us determine the best linear predictor for $\tilde{y}$ based upon $\phi(x)$. Let $\tilde{\mathbf{y}} = (\tilde{y}_1 \ldots \tilde{y}_n)'$. The least-squares prediction for $\tilde{y}$ is $\tilde{\mathbf{w}}'\phi(x)$ where

$$\begin{aligned}
\tilde{\mathbf{w}} &= (X'X)^{-1}X'\tilde{\mathbf{y}} && (14) \\
&= (X'X)^{-1}X'(\mathbf{y} - X\hat{\mathbf{w}}) && (15) \\
&= (\hat{\mathbf{w}} - (X'X)^{-1}(X'X)\hat{\mathbf{w}}) && (16) \\
&= (\hat{\mathbf{w}} - \hat{\mathbf{w}}) && (17) \\
&= \mathbf{0} && (18)
\end{aligned}$$

Hence, the best linear prediction of $\tilde{y}$ based upon $\phi(x)$ is $\mathbf{0}'\phi(x) = 0$ for all x.

(1-4) (5pts) Let $\tilde{\phi}(x) = A\phi(x)$ where $A = diag(a_1, \ldots, a_n)$ is an invertible $d \times d$ matrix ($a_i \neq 0$ for all i). The linear least-squares estimate of y based upon $\phi(x)$ is $\hat{\mathbf{w}}'\phi(x)$ with $\hat{\mathbf{w}} = (X'X)^{-1}X'\mathbf{y}$ where $X = (\phi(x_1) \ldots \phi(x_n))'$ and $\mathbf{y} = (y_1 \ldots y_n)'$. Similarly, the linear least-squares estimate of y based upon $\tilde{\phi}(x)$ is $\tilde{\mathbf{w}}'\tilde{\phi}(x)$ with $\tilde{\mathbf{w}} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\mathbf{y}$ where $\tilde{X}' = (\tilde{\phi}(x_1) \ldots \tilde{\phi}(x_n)) = AX'$. Then,

$$\begin{aligned}
\tilde{\mathbf{w}}'\tilde{\phi}(x) &= \{(AX'XA')^{-1}AX'\mathbf{y}\}'A\phi(x) && (19) \\
&= \{(A')^{-1}(X'X)^{-1}A^{-1}A\mathbf{y}\}'A\phi(x) && (20) \\
&= \{(X'X)^{-1}(A^{-1}A)\mathbf{y}\}'(A^{-1}A)\phi(x) && (21) \\
&= \{(X'X)^{-1}\mathbf{y}\}'\phi(x) && (22) \\
&= \hat{\mathbf{w}}'\phi(x) && (23)
\end{aligned}$$

(1-5) (Optional) Your MATLAB script should perform the following calculations:

For $\phi(x) = (1\ x\ x^2)'$;

$$X' = (\phi(x_1) \ldots \phi(x_6)) = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \\ 4 & 1 & 0 & 1 & 2 \end{pmatrix} \qquad (24)$$

$$K = X'X = \begin{pmatrix} 5 & 0 & 10 \\ 0 & 10 & 0 \\ 10 & 0 & 34 \end{pmatrix} \qquad (25)$$

$$b = X'y = \begin{pmatrix} -3 \\ 7 \\ -5 \end{pmatrix} \tag{26}$$

We then solve $Kw = b$ for the least squares parameters:

$$w = K^{-1}b \approx \begin{pmatrix} -0.74286 \\ 0.70000 \\ 0.07143 \end{pmatrix} \tag{27}$$

The prediction errors are

$$\mathbf{e} = \mathbf{y} - X\hat{\mathbf{w}} \approx \begin{pmatrix} -0.14286 \\ 0.37143 \\ -0.25714 \\ -0.02857 \\ 0.05714 \end{pmatrix} \tag{28}$$

In MATLAB it is easy to check that $\sum_i e_i = 0$ and $\sum_i e_i \phi(x_i) = 0$ (to within machine precision, `eps` $\approx 10^{-16}$).

The MATLAB script `hw1prob1.m` will perform these calculations and generate a plot:

```
% calculate least-squares params
x = [-2 -1 0 1 2]'
y = [-2 -1 -1 0 1]'
X = [ones(size(x)),x,x.^2]
K = X'*X b = X'*y
wh = K \ b % solves K w = b

% check that prediction error uncorrelated with features
yh = X*wh    % predictions
eh = y - yh  % prediction errors
me = mean(eh)
z = zeros(3,1);
for i=1:3
  z=z+eh(i)*X(i,:)';
end disp(z)

% generate plot
xx = [-3:.01:3]';
XX = [ones(size(xx)),xx,xx.^2];
yy = XX * wh;
plot(x,y,'o',xx,yy,'-');
```

For $\phi(x) = \sin \pi x$, the problem is ill-posed because for the given data $\phi(x_i) = 0$ for all x_i so that $\hat{y}_i = w \times 0 = 0$ for all i (no matter how we choose w). There is no basis for performing linear predictions of y values based on this feature function for the given data set. Note, however, that MATLAB does not necessarily evaluate $\sin \pi$ to be exactly zero but some small number. So, you would most likely get a clear answer to this problem (other than what you would expect) if you went ahead and solved it numerically in a straightforward manner. One needs to be a bit careful to avoid such numerical issues when implementing machine learning methods in practice.

Problem 2

(2-1) (10pts) Let $W = (\mathbf{w}_1\, \mathbf{w}_2)$ and $Y = (\mathbf{y}^1\, \mathbf{y}^2)$. The cost function may be decomposed into two parts;

$$\begin{aligned}
J(W;Y) &= \frac{1}{n}\sum_i \|\mathbf{y}_i - W'\phi(x_i)\|^2 && (29)\\
&= \frac{1}{n}\sum_i \left((y_{i,1} - \mathbf{w}_1'\phi(x_i))^2 + (y_{i,2} - \mathbf{w}_2'\phi(x_i))^2\right) && (30)\\
&= \left(\frac{1}{n}\sum_i (y_{i,1} - \mathbf{w}_1'\phi(x_i))^2\right) + \left(\frac{1}{2}\sum_i (y_{i,2} - \mathbf{w}_2'\phi(x_i))^2\right) && (31)\\
&= J_1(\mathbf{w}_1;\mathbf{y}^1) + J_2(\mathbf{w}_2;\mathbf{y}^2) && (32)
\end{aligned}$$

Hence, we choose $\hat{\mathbf{w}}_1$ s.t. $\hat{\mathbf{w}}_1'\phi(x)$ is the linear least-squares estimate of y_1 based upon $\phi(x)$. Likewise, $\hat{\mathbf{w}}_2$ is chosen s.t. $\hat{\mathbf{w}}_2'\phi(x)$ is the linear least-squares estimate of y_2 based upon $\phi(x)$. These least-squares parameters are given by:

$$\begin{aligned}
\hat{\mathbf{w}}_1 &= (X'X)^{-1}X'\mathbf{y}^1 && (33)\\
\hat{\mathbf{w}}_2 &= (X'X)^{-1}X'\mathbf{y}^2 && (34)
\end{aligned}$$

Concatenating column vectors yields:

$$\begin{aligned}
\hat{W} &= (\hat{\mathbf{w}}_1\, \hat{\mathbf{w}}_2) && (35)\\
&= (X'X)^{-1}X'(\mathbf{y}^1\, \mathbf{y}^2) && (36)\\
&= (X'X)^{-1}X'Y && (37)
\end{aligned}$$

(2-2) (5pts)

$$X' = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \quad (38)$$

$$X'X = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix} \quad (39)$$

$$Y = \begin{pmatrix} -1 & -1 \\ -1 & -2 \\ -2 & -1 \\ 1 & 1 \\ 1 & 2 \\ 2 & 1 \end{pmatrix} \quad (40)$$

$$X'Y = \begin{pmatrix} -4 & -4 \\ 4 & 4 \end{pmatrix} \quad (41)$$

$$\hat{W} = \frac{1}{3}I(X'Y) = \begin{pmatrix} -\frac{4}{3} & -\frac{4}{3} \\ \frac{4}{3} & \frac{4}{3} \end{pmatrix} \quad (42)$$

The MATLAB script `hw1prob2.m` will perform these calculations and generate a plot.

```
x = [0 0 0 1 1 1]'
X = [~x x]

Y = [-1 -1; -1 -2; -2 -1; 1 1; 1 2; 2 1]
```

```
Wh = inv(X'*X)*X'*Y
y1=Y(:,1)
y2=Y(:,2)
w1=Wh(:,1)
w2=Wh(:,2)
plot(y1,y2,'x',w1,w2,'o');
```

(2-3) (5pts)

$$\begin{aligned}
\sum_i \hat{\mathbf{e}}_i \phi(x_i) &= \left\{ \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \end{pmatrix} + \begin{pmatrix} \frac{1}{3} \\ -\frac{2}{3} \end{pmatrix} + \begin{pmatrix} -\frac{2}{3} \\ \frac{1}{3} \end{pmatrix} \right\} (1\ 0) + \left\{ \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \end{pmatrix} + \begin{pmatrix} \frac{1}{3} \\ -\frac{2}{3} \end{pmatrix} + \begin{pmatrix} -\frac{2}{3} \\ \frac{1}{3} \end{pmatrix} \right\} (0\ 1) \\
&= \begin{pmatrix} 0 \\ 0 \end{pmatrix} (1\ 0) + \begin{pmatrix} 0 \\ 0 \end{pmatrix} (0\ 1) && (43) \\
&= \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} && (44)
\end{aligned}$$

Note that $\hat{\mathbf{y}}_0 = \mathbf{f}(0; \hat{W}) = (-\frac{4}{3}\ -\frac{4}{3})'$ is the (conditional) sample average of $\mathbf{y}$ over those samples where $x = 0$. Likewise, $\hat{\mathbf{y}}_1 = \mathbf{f}(1, \hat{W}) = (\frac{4}{3}\ \frac{4}{3})'$ is the sample average of $\mathbf{y}$ over those samples where $x = 1$. Consequently, $\sum_i \hat{\mathbf{e}}_i = 0$.

Part 2: Probabilistic Modelling and Likelihood

(No Problem 3)

Problem 4

The pmf of $x \in \{0, 1\}$ is

$$P(x) = \begin{cases} 1 - \theta_1, & x = 0 \\ \theta_1, & x = 1 \end{cases} \tag{45}$$

The conditional pmf of $y \in \{0, 1\}$ given that $x = 0$ is

$$P(y|x = 0) = \begin{cases} \theta_2, & y = 0 \\ 1 - \theta_2, & y = 1 \end{cases} \tag{46}$$

The conditional pmf of y given that $x = 1$ is

$$P(y|x = 1) = \begin{cases} 1 - \theta_2, & y = 0 \\ \theta_2, & y = 1 \end{cases} \tag{47}$$

(4-1) (5pts) Use $P(x, y) = P(y|x)P(x)$ to tabulate the joint pmf of (x, y).

$$P_{\mathrm{x,y}} \equiv \begin{pmatrix} P(0,0) & P(0,1) \\ P(1,0) & P(1,1) \end{pmatrix} = \begin{pmatrix} \theta_2(1-\theta_1) & (1-\theta_2)(1-\theta_1) \\ (1-\theta_2)\theta_1 & \theta_2\theta_1 \end{pmatrix} \tag{48}$$

(4-2) (10pts) We select (θ_1, θ_2) to minimize the log-likelihood of the samples $\{(x_i, y_i), i = 1, ..., n\}$ which may be expressed as

$$J(\theta_1, \theta_2) = \sum_i \log P(x_i, y_i) \tag{49}$$

$$
\begin{aligned}
&= \sum_i (\log P(y_i|x_i) + \log P(x_i)) && (50)\\
&= \left(\sum_i \log P(y_i|x_i)\right) + \left(\sum_i \log P(x_i)\right) && (51)\\
&= J_2(\theta_2) + J_1(\theta_1) && (52)
\end{aligned}
$$

Hence, we choose θ_1 to minimize

$$
\begin{aligned}
J_1(\theta_1) &= \sum_i \log P(x_i) && (53)\\
&= N(x=1)\log\theta_1 + (n - N(x=1))\log(1-\theta_1) && (54)
\end{aligned}
$$

where $N(x = 1) = \sum_i x_i$. Differentiating w.r.t. θ_1 gives

$$
\frac{\partial J_1}{\partial \theta_1} = \frac{N(x=1)}{\theta_1} - \frac{n - N(x=1)}{1-\theta_1} \tag{55}
$$

We set this derivative to zero and solve for θ_1 to obtain

$$
\hat{\theta}_1 = \frac{N(x=1)}{n} \tag{56}
$$

Similarly, we choose θ_2 to minimize

$$
\begin{aligned}
J_2(\theta_2) &= \sum_i \log P(y_i|x_i) && (57)\\
&= N(x=y)\theta_2 + (n - N(x=y))(1-\theta_2) && (58)\\
& && (59)
\end{aligned}
$$

where $N(x = y) = \sum_i (x_i y_i + (1 - x_i)(1 - y_i))$. Differentiating J_2 w.r.t. θ_2, setting to zero and solving for θ_2 gives

$$
\hat{\theta}_2 = \frac{N(x=y)}{n} \tag{60}
$$

For the example data;

$$
\begin{aligned}
\hat{\theta}_1 &= \frac{4}{7} && (61)\\
\hat{\theta}_2 &= \frac{4}{7} && (62)
\end{aligned}
$$

The maximum likelihood of the data under this model is

$$
\prod_i \hat{P}(y_i|x_i)\hat{P}(x_i) = \left(\frac{4}{7}\right)^8 \left(\frac{3}{7}\right)^6 \approx 7.0443 \times 10^{-5} \tag{63}
$$

(4-3) (10pts) The expected value of the estimate $\hat{\theta}_1 = \frac{1}{n}\sum_i \mathrm{x}_i$ is

$$
\begin{aligned}
E\{\hat{\theta}_1(\mathrm{x}_1, \ldots, \mathrm{x}_2)\} &= E\left\{\frac{1}{n}\sum_i \mathrm{x}_i\right\} && (64)\\
&= \frac{1}{n}\sum_i E\{\mathrm{x}_i\} && (65)\\
&= \frac{1}{n}\sum_i \theta_1 && (66)\\
&= \theta_1 && (67)
\end{aligned}
$$

Hence, the ML estimate $\hat{\theta}_1$ is unbiased.

(4-4) (10pts) There are four possible outcomes $(x, y) \in \{(0,0),(0,1),(1,0),(1,1)\}$. Label these outcomes $z = 0, 1, 2, 3$. A minimal parameterization of the joint pmf $P(z)$ is given below:

$$\begin{aligned} P(0) &= 1 - \sum_{k=1}^{3} \theta_k && (68) \\ P(1) &= \theta_1 && (69) \\ P(2) &= \theta_2 && (70) \\ P(3) &= \theta_3 && (71) \end{aligned}$$

Given the samples $\{(x_i, y_i), i = 1, \ldots, n\}$ the log-likelihood is

$$\begin{aligned} J(\theta) &= \sum_i \log P(x_i, y_i) && (72) \\ &= N_0 \log(1 - \sum_{i=1}^{3} \theta_i) + \sum_{j=1}^{3} N_j \log \theta_j && (73) \end{aligned}$$

where N_k is the number of times $z = k$ occurs in the observed samples. Differentiating w.r.t. to each θ_k gives

$$\frac{\partial J}{\partial \theta_k} = -\frac{N_0}{1 - \sum_{i=1}^{3} \theta_i} + \frac{N_k}{\theta_k} \tag{74}$$

Setting each derivative to zero, we obtain $\theta_k = \frac{N_k}{\lambda}$ where $\lambda = N_0/(1 - \sum_{i=1}^{3} \theta_i)$; substitution gives $\lambda = N_0/(1 - \frac{1}{\lambda}(N_1 + N_2 + N_3))$; solve for $\lambda = N_0 + N_1 + N_2 + N_3 = n$. Hence, the ML estimate of the pmf of z is $P(\mathrm{z} = k) = \frac{N_k}{n}$. Equivalently, the ML estimate of joint pmf of (x, y) is

$$\hat{P}(x, y) = \frac{N(x, y)}{n} \tag{75}$$

where $N(x, y)$ is the number of times (x, y) occurs in the observed samples.

For the example data;

$$\hat{P}_{\mathrm{x,y}} \equiv \begin{pmatrix} \hat{P}(0,0) & \hat{P}(0,1) \\ \hat{P}(1,0) & \hat{P}(1,1) \end{pmatrix} = \begin{pmatrix} \frac{2}{7} & \frac{1}{7} \\ \frac{2}{7} & \frac{2}{7} \end{pmatrix} \tag{76}$$

The maximum likelihood of the data under this model is

$$\prod_i \hat{P}(x_i, y_i) = \left(\frac{1}{7}\right)^1 \left(\frac{2}{7}\right)^6 = \frac{64}{823543} \approx 7.7712 \times 10^{-5} \tag{77}$$

which is higher than in the previous two-parameter model (as we would expect since the two-parameter model is contained by the three-parameter model).

(4-5) (Optional) Let $\delta(u, v)$ be defined so that $\delta(u, v) = 1$ if $u = v$ and $\delta(u, v) = 0$ otherwise. Then, $N(x, y) = \sum_i \delta(x, x_i)\delta(y, y_i)$ and

$$\begin{aligned} E\{\hat{P}(x, y)\} &= \frac{1}{n} \sum_i E\{\delta(x, \mathrm{x}_i)\delta(y, \mathrm{y}_i)\} && (78) \\ &= \frac{1}{n} \sum_i P(x, y) && (79) \\ &= P(x, y) && (80) \end{aligned}$$

(4-6) (Optional) Let $\hat{\theta}^{\setminus i}$ denote the ML estimate of the model parameters based upon samples $\{1, \ldots, n\} \setminus i$ (omitting sample i). This generates n estimates of the model parameters θ. For the i-th estimate, compute the log-likelihood of sample i. Sum this leave-one-out log-likelihood statistic over all samples.

$$J = \sum_i \log P(x_i, y_i; \hat{\theta}^{\setminus i}) \tag{81}$$

Compute this cross-validation log-likelihood under both models and prefer the model which produces the higher value.

For the two-parameter model we calculate

$$\begin{aligned} J &= \log\left(\frac{1}{4}\cdot\frac{1}{6}\cdot\frac{1}{6}\cdot\frac{1}{6}\cdot\frac{1}{4}\cdot\frac{1}{3}\cdot\frac{4}{9}\right) && (82)\\ &= \log\frac{1}{23328} && (83)\\ &\approx -10.057 && (84)\end{aligned}$$

where we have taken log to be the natural logarithm.

For the three-parameter model, note that the last sample $(x_7, y_7) = (0, 1)$ is the only occurrence of $(0, 1)$ in the data set. Hence, $\hat{P}^{\setminus 7}(0, 1) = 0$ and $J = \log 0 = -\infty$. This suggests that the three-parameter model has overfit the data and we should favor the two-parameter model.

Problem 5

(5-1) (10pts) We wish to maximize the log-likelihood of observed samples $\{\mathbf{x}_i, i = 1, \ldots, n\}$.

$$\begin{aligned} L(\mu, \Sigma) &= \sum_i \log p(\mathbf{x}_i; \mu, \Sigma) && (85)\\ &= -\frac{1}{2}\{n \log|\Sigma| + \sum_i (\mathbf{x}_i - \mu)'\Sigma^{-1}(\mathbf{x}_i - \mu)\} + \text{const} && (86)\end{aligned}$$

Calculate the derivative of L w.r.t. the mean parameters μ and the inverse-covariance parameters $A = \Sigma^{-1}$:

$$\begin{aligned} \frac{dL}{d\mu} &= -\frac{1}{2}\sum_i \frac{d}{d\mu}\left\{(\mathbf{x}_i - \mu)'\Sigma^{-1}(\mathbf{x}_i - \mu)\right\} && (87)\\ &= -\frac{1}{2}\sum_i 2\Sigma^{-1}(\mathbf{x}_i - \mu) && (88)\\ &= n\Sigma^{-1}\left(\mu - \frac{1}{n}\sum_i \mathbf{x}_i\right) && (89)\end{aligned}$$

$$\begin{aligned} \frac{dL}{dA} &= -\frac{1}{2}\left\{-n\frac{d\log|A|}{dA} + \sum_i \frac{d}{dA}(\mathbf{x}_i - \mu)'A(\mathbf{x}_i - \mu)\right\} && (90)\\ &= \frac{n}{2}\left\{A^{-1} - \frac{1}{n}\sum_i (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)'\right\} && (91)\end{aligned}$$

Solving the system of equations

$$\begin{aligned} \frac{dL}{du} &= 0 && (92)\\ \frac{dL}{dA} &= 0 && (93)\end{aligned}$$

for $(\mu, \Sigma = A^{-1})$ gives the joint ML estimates:

$$\hat{\mu} = \frac{1}{n} \sum_i \mathbf{x}_i \tag{94}$$

$$\hat{\Sigma} = \frac{1}{n} \sum_i (\mathbf{x}_i - \hat{\mu}) \tag{95}$$

(5-2) There are several possible ways of solving this problem. We will proceed here in a way that explicates some useful properties of Gaussian distributions. Let

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \tag{96}$$

and

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{pmatrix} \tag{97}$$

Note that $\mu_i = E\{\mathrm{x}_i\}$, $\sigma_i^2 = \mathrm{var(x_i)}$ and $\sigma_{1,2} = \mathrm{cov}(\mathrm{x}_1, \mathrm{x}_2)$. The marginal distributions of a bivariate Gaussian distribution are univariate Gaussian distributions (a well known fact which you do not have to prove). Hence, $\mathrm{x}_1 \sim N(\mu_1, \sigma_1^2)$ and the pdf $p(x_1)$ is

$$p(x_1) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left\{ -\frac{1}{2} \left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 \right\} \tag{98}$$

The conditional pdf's of bivariate Gaussian are also (conditional) univariate Gaussian distributions. We explicitly show this thereby determining $E\{\mathrm{x}_2 | \mathrm{x}_1 = x_1\}$. First, note that the inverse covariance is

$$\Sigma^{-1} = \frac{1}{|\Sigma|} \begin{pmatrix} \sigma_2^2 & -\sigma_{1,2} \\ -\sigma_{1,2} & \sigma_1^2 \end{pmatrix} \tag{99}$$

$$= \frac{1}{1-\rho^2} \begin{pmatrix} \frac{1}{\sigma_1^2} & -\frac{\rho}{\sigma_1\sigma_2} \\ -\frac{\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{pmatrix} \tag{100}$$

where $|\Sigma| = \sigma_1^2\sigma_2^2 - \sigma_{1,2}^2 = \sigma_1^2\sigma_2^2(1-\rho^2)$ and $\rho = \frac{\sigma_{1,2}}{\sigma_1\sigma_2}$. Write out the joint pdf in (x_1, x_2).

$$p(x_1, x_2) = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left(\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1} \right) \left(\frac{x_2 - \mu_2}{\sigma_2} \right) \right) \right\} \tag{101}$$

The conditional pdf of x_2 given x_1 (up to the normalization constant) is

$$\begin{aligned} p(x_2|x_1) &= \frac{p(x_1, x_2)}{p(x_1)} \\ &\propto \exp\left\{ -\frac{1}{2(1-\rho^2)} \left(\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1} \right) \left(\frac{x_2 - \mu_2}{\sigma_2} \right) \right) + \frac{1}{2} \left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 \right\} \\ &\propto \exp\left\{ -\frac{1}{2(1-\rho^2)} \left(\left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 + \rho^2 \left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1} \right) \left(\frac{x_2 - \mu_2}{\sigma_2} \right) \right) \right\} \\ &\propto \exp\left\{ -\frac{1}{2(1-\rho^2)} \left(\left(\frac{x_2 - \mu_2}{\sigma_2} \right) - \rho \left(\frac{x_1 - \mu_1}{\sigma_1} \right) \right)^2 \right\} \end{aligned}$$

$$\propto \quad \exp\left\{-\frac{1}{2}\left(\frac{x_2 - \left(\mu_2 + \sigma_2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\right)}{\sigma_2\sqrt{1-\rho^2}}\right)^2\right\}$$

$$\propto \quad \exp\left\{-\frac{1}{2}\left(\frac{x_2 - \mu_{2|1}(x_1)}{\sigma_{2|1}}\right)^2\right\} \tag{102}$$

where

$$\mu_{2|1}(x_1) = \mu_2 + \frac{\sigma_2\rho}{\sigma_1}(x_1 - \mu_1) \tag{103}$$

$$\sigma_{2|1}^2 = \sigma_2^2(1-\rho^2) \tag{104}$$

This shows that the conditional distribution of x_2 given x_1 is the univariate Gaussian distribution $N(\mu_{2|1}(x_1), \sigma_{2|1}^2)$ with (conditional) mean $E\{\mathrm{x}_2|x_1\} = \mu_{2|1}(x_1)$ and (conditional) variance $\mathrm{var}(\mathrm{x}_2|x_1) = \sigma_{2|1}^2$.

Hence, the minimum mean-square error (MMSE) estimate of x_2 given x_1 is

$$\hat{x}_2(x_1) = \mu_2 + \frac{\sigma_2\rho}{\sigma_1}(x_1 - \mu_1) \tag{105}$$

$$= \mu_2 + \frac{\sigma_{1,2}}{\sigma_1^2}(x_1 - \mu_1) \tag{106}$$

which is what we were asked to derive. Note that this estimate happens to be linear in x_1 (although we did not require this) and hence agrees with the formula for the linear least-squares estimate of x_2 based upon x_1 (derived in recitation). In general, for jointly Gaussian random variables, linear least-squares estimation is equivalent to minimum mean-square estimation.