ELE 535: Machine Learning and Pattern Recognition
Homework 6

Zachary Hervieux-Moore

Wednesday 14th November, 2018

Exercise 1: The density $f_X(x)$ takes the exponential form

$$f_X(x) = \frac{h(x)}{Z(\theta_0)} e^{\langle \theta_0, t(x) \rangle}$$

You want to use data from m i.i.d. draws from f_X to estimate the value of θ_0. Given the training data, the likelihood function for any parameter value θ is

$$L(\theta) = \prod_{i=1}^{m} \frac{1}{Z(\theta)} h(x_i) e^{\langle \theta, t(x_i) \rangle}$$

The maximum likelihood estimate $\hat{\theta}_0$ of θ_0 is obtained by solving

$$\hat{\theta}_0 = \arg\max_{\theta} L(\theta)$$

Show that $\hat{\theta}_0$ must satisfy

$$\nabla \ln(Z(\hat{\theta}_0)) = \frac{1}{m} \sum_{i=1}^{m} t(x_i)$$

Answer: Taking the log of the likelihood,

$$\ln(L(\theta)) = \sum_{i=1}^{m} \ln(h(x_i)) + \langle \theta, t(x_i) \rangle - \ln(Z(\theta))$$

Now, taking the derivative and setting it to 0,

$$\nabla_\theta \ln(L(\theta)) = \sum_{i=1}^{m} t(x_i) - \nabla \ln(Z(\theta)) = 0$$

$$\implies \nabla \ln(Z(\theta)) = \frac{1}{m} \sum_{i=1}^{m} t(x_i)$$

Because the log likelihood of exponential families is concave, the $\hat{\theta}_0$ that satisfies this equation is the maximizer and we conclude that this shows the result.

Exercise 2: Given m i.i.d. draws from a multivariate Gaussian density, use the method in (Q1) to find the maximum likelihood estimates of the mean μ and covariance matrix Σ of the density.

Answer: To use the result above, we parameterize the multivariate Gaussian as:

$$\theta = (\Sigma^{-1}\mu, \Sigma^{-1}) = (\theta_1, \theta_2) \quad h(x) = 1 \quad t(x) = (x, -\frac{1}{2}xx^T)$$
$$Z(\theta) = (2\pi)^{\frac{n}{2}} det(\theta_2)^{-\frac{1}{2}} e^{\frac{1}{2}\theta_1^T\theta_2^{-1}\theta_1}$$

Now, plugging this all into the above, first for θ_1,

$$\nabla_{\theta_1} \ln(Z(\theta)) = \nabla_{\theta_1}\left[\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln(det(\theta_2)) + \frac{1}{2}\theta_1^T\theta_2^{-1}\theta_1\right] = \frac{1}{m}\sum_{i=1}^{m} x_i$$
$$\implies \theta_2^{-1}\theta_1 = \frac{1}{m}\sum_{i=1}^{m} x_i$$
$$\implies \Sigma\Sigma^{-1}\mu = \frac{1}{m}\sum_{i=1}^{m} x_i$$
$$\implies \mu = \frac{1}{m}\sum_{i=1}^{m} x_i$$

Now for θ_2,

$$\nabla_{\theta_2} \ln(Z(\theta)) = \nabla_{\theta_2}\left[\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln(det(\theta_2)) + \frac{1}{2}\theta_1^T\theta_2^{-1}\theta_1\right] = \frac{1}{m}\sum_{i=1}^{m} -\frac{1}{2}x_ix_i^T$$
$$\implies -\frac{1}{2}\theta_2^{-T} + \frac{1}{2}\theta_2^{-T}\theta_1\theta_1^T\theta_2^{-T} = \frac{1}{m}\sum_{i=1}^{m} -\frac{1}{2}x_ix_i^T$$
$$\implies \Sigma^T + \Sigma^T\Sigma^{-1}\mu\mu^T\Sigma^{-T}\Sigma^T = \frac{1}{m}\sum_{i=1}^{m} x_ix_i^T$$

Using the fact that Σ is symmetric and the definition of μ above,

$$\Sigma + \mu\mu^T = \frac{1}{m}\sum_{i=1}^{m} x_i x_i^T$$

$$\implies \Sigma = \frac{1}{m}\sum_{i=1}^{m}(x_i x_i^T + \mu\mu^T) - 2\mu\mu^T$$

$$\implies \Sigma = \frac{1}{m}\sum_{i=1}^{m}(x_i x_i^T + \mu\mu^T) - \mu\left(\frac{1}{m}\sum_{i=1}^{m} x_i\right)^T - \left(\frac{1}{m}\sum_{i=1}^{m} x_i\right)\mu^T$$

$$\implies \Sigma = \frac{1}{m}\sum_{i=1}^{m}(x_i x_i^T - \mu x_i^T - x_i\mu^T + \mu\mu^T)$$

$$\implies \Sigma = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu)(x_i - \mu)^T$$

Thus, we have

$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\hat{\Sigma} = \frac{1}{m}\sum_{i=1}^{m}(x_i - \hat{\mu})(x_i - \hat{\mu})^T$$

Exercise 3: Empirical statistics, MSE affine prediction, and least squares. Fix a training dataset $\{(x_i, y_i\}_{i=1}^m$, with examples $x_i \in \mathbb{R}^n$ and targets $y_i \in \mathbb{R}^q$. Let X denote the matrix with the examples as its columns, Y denote the matrix with the corresponding targets as its columns. Define the following first and second order empirical statistics of the data:

$$\hat{\mu}_X = \frac{1}{m} X\mathbf{1}_m \quad \hat{\mu}_Y = \frac{1}{m} Y\mathbf{1}_m \tag{1}$$
$$\hat{\Sigma}_X = \frac{1}{m}(X - \hat{\mu}_X\mathbf{1}_m^T)(X - \hat{\mu}_X\mathbf{1}_m^T)^T \quad \hat{\Sigma}_{XY} = \frac{1}{m}(X - \hat{\mu}_X\mathbf{1}_m^T)(Y - \hat{\mu}_Y\mathbf{1}_m^T)^T$$

An optimal MSE affine estimator $\hat{y}(x) = W^T x + b$ based on the empirical statistics in (1) must satisfy

$$\hat{\Sigma}_X W = \hat{\Sigma}_{XY} \quad b = \hat{\mu}_Y - W^T \hat{\mu}_X \tag{2}$$

a) Consider the least squares problem

$$W_*, b_* = \underset{W\in\mathbb{R}^{n\times q}, b\in\mathbb{R}^q}{\arg\min} \|Y - W^T X - b\mathbf{1}_m^T\|_F^2 \tag{3}$$

Show that W_*, b_* satisfies (2). Thus directly solving the least squares problem (3) yields an optimal MSE affine estimator for the empirical first and second order statistics in (1).

b) Consider the ridge regression problem

$$W_{r*}, b_{r*} = \underset{W\in\mathbb{R}^{n\times q}, b\in\mathbb{R}^q}{\arg\min} \frac{1}{m}\|Y - W^T X - b\mathbf{1}_m^T\|_F^2 + \lambda\|W\|_F^2, \quad \lambda > 0$$

Determine if W_{r*}, b_{r*} satisfy (2). If not, what needs to be changd in (1) to ensure W_{r*}, b_{r*} satisfy (2). Interpret the change you suggest.

Answer:

a) Rewriting the optimization problem and checking first order neccessary conditions for b^*,

$$
\begin{gathered}
\|Y - W^T X - b\mathbf{1}_m^T\|_F^2 \\
\text{trace}((Y - W^T X - b\mathbf{1}_m^T)^T (Y - W^T X - b\mathbf{1}_m^T)) \\
\nabla_b : -Y\mathbf{1}_m + W^T X\mathbf{1}_m - Y\mathbf{1}_m + W^T X\mathbf{1}_m + 2b\mathbf{1}_m^T\mathbf{1}_m = 0 \\
\implies mb^* = (Y\mathbf{1}_m - W^{*T} X\mathbf{1}_m) \\
\implies b^* = \hat{\mu}_Y - W^{*T}\hat{\mu}_X
\end{gathered}
$$

Doing the same process for W^*,

$$
\begin{gathered}
\|Y - W^T X - b\mathbf{1}_m^T\|_F^2 \\
\text{trace}((Y - W^T X - b\mathbf{1}_m^T)^T (Y - W^T X - b\mathbf{1}_m^T)) \\
\nabla_W : -XY^T - XY^T + 2XX^T W^* + X\mathbf{1}b^{*T} + X\mathbf{1}_m b^{*T} = 0 \\
\implies XX^T W^* = XY^T - X\mathbf{1}_m b^{*T} \\
\implies XX^T W^* = XY^T - X\mathbf{1}_m(\hat{\mu}_Y - W^{*T}\hat{\mu}_X)^T \\
\implies (XX^T - X\mathbf{1}_m\hat{\mu}_X^T)W^* = XY^T - X\mathbf{1}_m\hat{\mu}_Y^T
\end{gathered}
$$

Adding and subtracting $\hat{\mu}_X\mathbf{1}_m^T\mathbf{1}_m\hat{\mu}_X^T$ and using the fact that $\mathbf{1}_m^T\mathbf{1}_m\hat{\mu}_X = X\mathbf{1}_m$ on the LHS. Doing the same thing on the RHS but with $\hat{\mu}_Y^T$ yields,

$$
\begin{gathered}
(XX^T - X\mathbf{1}_m\hat{\mu}_X^T - \hat{\mu}_X\mathbf{1}_m^T X + \hat{\mu}_X\mathbf{1}_m^T\mathbf{1}_m\hat{\mu}_X^T)W^* \\
= (XY^T - X\mathbf{1}_m\hat{\mu}_Y^T - \hat{\mu}_X\mathbf{1}_m^T Y + \hat{\mu}_X\mathbf{1}_m^T\mathbf{1}_m\hat{\mu}_Y^T) \\
\implies (X - \hat{\mu}_X\mathbf{1}_m^T)(X - \hat{\mu}_X\mathbf{1}_m^T)^T W^* = (X - \hat{\mu}_X\mathbf{1}_m^T)(Y - \hat{\mu}_Y\mathbf{1}_m^T)^T
\end{gathered}
$$

Dividing both sides by m yields the result,

$$\hat{\Sigma}_X W^* = \hat{\Sigma}_{XY}$$

b) The first thing to notice is that the new term in the ridge regression only involves W and this the first order neccessary conditions for b^* remains unchanged. Thus, b^* satisfies its original equation unchanged. Repeating the same steps as before, we get the following relations

$$(XX^T - X\mathbf{1}_m\hat{\mu}_X^T + \lambda I_n)W^* = XY^T - X\mathbf{1}_m\hat{\mu}_Y^T$$

Using the same trick as before yields,

$$(\hat{\Sigma}_X + \lambda I_n)W^* = \hat{\Sigma}_{XY}$$

Thus, we have to change $\hat{\Sigma}_X$ in (1) as follows:

$$\hat{\Sigma}_{X_{ridge}} = \frac{1}{m}(X - \hat{\mu}_X \mathbf{1}_m^T)(X - \hat{\mu}_X \mathbf{1}_m^T)^T + \frac{\lambda}{m} I_n$$

My interpretation of this is as follows. We know that MSE can be decomposed into a bias-variance tradeoff. By using an unbiased estimator, all of your error is a result of the variance of your estimator. Introducing the penalty term in ridge regression introduces bias; namely in the Σ_X term. This allows for control of the bias through λ which may reduce MSE of your estimator.

Exercise 4: Bias, error covariance, and MSE. Consider random vectors X and Y with a joint density f_{XY} and PD covariance Σ. Let X have mean $\mu_X \in \mathbb{R}^n$ and covariance $\Sigma_X \in \mathbb{R}^{n\times n}$, Y have $\mu_Y \in \mathbb{R}^q$ and covariance $\Sigma_Y \in \mathbb{R}^{q\times q}$, and let the cross-covariance of X and Y be $\Sigma_{XY} \in \mathbb{R}^{n\times q}$.

Let $\hat{y}(x)$ be an estimator of Y given $X = x$, and denote the corresponding prediction error by $E \triangleq Y - \hat{y}(X)$. Of interest is μ_E, Σ_E and the MSE. The estimator is said to be *unbiased* if $\mu_E = \mathbf{0}$.

a) For any estimator $\hat{y}$ with finite μ_E and MSE, show that $\text{MSE}(\hat{y}) = \text{trace}(\Sigma_E) + \|\mu_E\|_2^2$. This shows that the MSE is the sum of two terms: the total variance $\text{trace}(\Sigma_E)$ of the error, and the squared norm of the bias $\|\mu_E\|_2^2$.

b) Let $\hat{y}(x) = \mu_Y$. Show that this is an unbiased estimator, determine Σ_E, show that Σ_E is PD, and determine the estimator MSE.

c) The MMSE affine estimator of Y given $X = x$ is

$$\hat{y}^*(x) = \mu_Y + W^{*T}(x - \mu_X) \quad \text{with } \Sigma_X W^* = \Sigma_{XY}$$

Show that $\hat{y}^*(\cdot)$ is an unbiased estimator, determine Σ_E, show that Σ_E is PD, and determine the estimator MSE.

Answer:

a) We have that

$$\begin{aligned}
MSE(\hat{y}) &= \mathbb{E}[\|Y - \hat{y}\|_2^2] = \mathbb{E}[(Y - \hat{y})^T(Y - \hat{y})] \\
&= \mathbb{E}[(Y - \hat{y} - \mu_E + \mu_E)^T(Y - \hat{y} - \mu_E + \mu_E)] \\
&= \mathbb{E}[(Y - \hat{y} - \mu_E)^T(Y - \hat{y} - \mu_E)] + \mathbb{E}[\mu_E^T\mu_E] \\
&\quad + \mathbb{E}[\mu_E^T(Y - \hat{y} - \mu_E)] + \mathbb{E}[(Y - \hat{y} - \mu_E)^T\mu_E]
\end{aligned}$$

Expanding the above and using the fact that everything is a scalar and μ_E is a constant,

$$\begin{aligned}
&= \mathbb{E}[trace((Y - \hat{y} - \mu_E)^T(Y - \hat{y} - \mu_E))] + \|\mu_E\|_2^2 \\
&\quad -2\mu_E^T\mathbb{E}[\hat{y}] - \|\mu_E\|_2^2 + 2\mu_E^T\mu_Y
\end{aligned}$$

Now we use the following facts: $\mathbb{E}[\hat{y}] = \mu_Y - \mu_E$, the cyclic property of trace, and the fact that trace is linear so we exchange the expectation and the trace,

$$\begin{aligned} &= trace(\mathbb{E}[(Y - \hat{y} - \mu_E)(Y - \hat{y} - \mu_E)^T]) + \|\mu_E\|_2^2 \\ &= trace(\mathbb{E}[(E - \mu_E)(E - \mu_E)^T]) + \|\mu_E\|_2^2 \\ &= trace(\Sigma_E) + \|\mu_E\|_2^2 \end{aligned}$$

b) If $\hat{y}(x) = \mu_Y$ then $E = Y - \mu_Y$. This also means that

$$\mu_E = \mathbb{E}[E] = \mathbb{E}[Y - \mu_Y] = \mathbb{E}[Y] - \mu_Y = \mu_Y - \mu_Y = 0$$

Thus, we have that it is unbiased. Now, we then have the following expression for Σ_E

$$\Sigma_E = \mathbb{E}[(E - \mu_E)(E - \mu_E)^T] = \mathbb{E}[(Y - \mu_Y)(Y - \mu_Y)^T] = \Sigma_Y$$

As Σ_Y is PD (Σ is PD implies Σ_Y is PD by picking vectors that zero out the X components), so is Σ_E. Using part a), we get that the MSE is:

$$MSE(\hat{y}) = trace(\Sigma_Y)$$

c) We check if it is unbiased first,

$$\begin{aligned} \mu_E = \mathbb{E}[E] = \mathbb{E}[Y - \hat{y}] &= \mu_Y - \mathbb{E}[\mu_Y + {W^*}^T(x - \mu_X)] \\ &= -{W^*}^T\mathbb{E}[x] + {W^*}^T\mu_X = 0 \end{aligned}$$

Thus, it is unbiased. Now, by definition

$$\begin{aligned} \Sigma_E &= \mathbb{E}[(Y - \mu_Y - {W^*}^T(x - \mu_X))(Y - \mu_Y - {W^*}^T(x - \mu_X))^T] \\ &= \mathbb{E}[(Y - \mu_Y)(Y - \mu_Y)^T] - \mathbb{E}[(Y - \mu_Y)(x - \mu_X)^T]W^* \\ &\quad -{W^*}^T\mathbb{E}[(x - \mu_X)(Y - \mu_Y)^T] + {W^*}^T\mathbb{E}[(x - \mu_X)(x - \mu_X)^T]W^* \\ &= \Sigma_Y - \Sigma_{YX}W^* - {W^*}^T\Sigma_{XY} + {W^*}^T\Sigma_X W^* \end{aligned}$$

Using our relation that $\Sigma_X W^* = \Sigma_{XY}$ and $\Sigma_{XY} = \Sigma_{YX}^T$,

$$\begin{aligned} &= \Sigma_Y - 2{W^*}^T\Sigma_X W^* + {W^*}^T\Sigma_X W^* \\ &= \Sigma_Y - {W^*}^T\Sigma_X W^* \end{aligned}$$

This is PD by the follwing Schur complement fact:

$$\Sigma = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix} \succ 0 \iff \Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY} \succ 0$$

Using the condition $\Sigma_X W^* = \Sigma_{XY}$ we get that

$$\Sigma_E = \Sigma_Y - {W^*}^T \Sigma_X W^* \succ 0$$

Thus, we have that Σ_E is PD and part a) shows that the MSE is

$$\begin{gathered} MSE(\hat{y}) = trace(\Sigma_E) = trace(\Sigma_Y) - trace({W^*}^T \Sigma_X W^*) \\ \text{or } trace(\Sigma_Y) - trace(\Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}) \end{gathered}$$

Exercise 5: The derivative and gradient of $\|M\|_2$. For $M \in \mathbb{R}^{m\times n}$ define $f(M) = \|M\|_2$. Determine a sufficient condition for the derivative of f to exist at M, and under these conditions find $Df(M)(H)$ and $\nabla f(M)$.

Answer: By SVD, we have that $M = U\Sigma V^T$. We also know that $\|M\|_2 = \sigma_1(M) = \Sigma_{11}$. Rearranging the SVD yields $\Sigma = U^T M \Sigma V$. Writing this out in summation form,

$$\Sigma_{11} = \sum_{j=1}^{n}\sum_{i=1}^{m} u_{1_i} M_{ij} v_{1_j}$$

Where u_1 and v_1 are the first column vector of U and V. Thus, we have that

$$\frac{\partial \Sigma_{11}}{\partial M_{ij}} = u_{1_i} v_{1_j} \text{ and so } Df(M)(H) = \sum_{j=1}^{n}\sum_{i=1}^{m} u_{1_i} v_{1_j} H_{ij}$$

So we conclude that

$$\nabla f(M) = u_1 v_1^T \text{ and } Df(M)(H) = trace(u_1 v_1^T H^T)$$

We note here that the analysis above only follows if $\sigma_1 > \sigma_2$. If $\sigma_1 = \sigma_2$, then one must change the analysis to be subgradients as the largest singular value would have competing derivatives from $u_1 v_1^T$ and $u_2 v_2^T$.