STA414/2104: Partial Problem Set Solutions, Problems 1 - 3

January, 2018

Problems 1, Question 1: Let $\mathbf{x} = (X_1, \ldots, X_j)$ be a vector-valued random variable taking values in the vector space $\mathbb{R}^p$. Define the *covariance matrix* of $\mathbf{x}$ to be $\Sigma = E((\mathbf{x} - E(\mathbf{x})(\mathbf{x} - E(\mathbf{x}))^T)$, which satisfies $\Sigma_{ij} = Cov(X_i, X_j)$. E here means *expectation*, not *error*.

(a) Show $Var(\mathbf{a}^T\mathbf{x}) = \mathbf{a}^T\Sigma\mathbf{a}$ for any fixed $\mathbf{a} \in \mathbb{R}^p$.

(b) Prove that Σ is positive definite. You may assume that $\mathbf{x}$ is what we call a *regular* random variable, which just means that every element of $\mathbf{x}$ has variance that is strictly positive. This question is a one-liner if you use the *definition* of positive definite-ness, so really what I'm asking you to do is look up the definition of "positive-definite" and clearly state why a covariance matrix should be so, given part (a).

Solution:

(a)

$$
\begin{aligned}
Var(\mathbf{a}^T\mathbf{x}) &= Var\left(\sum_{i=1}^{n} a_i X_i\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} Var(a_i X_i) + Var(a_j X_j) + 2Cov(a_i X_i, a_j X_j) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i^2 Var(X_i) + a_j^2 Var(X_j) + 2a_i a_j Cov(X_i, X_j) \\
&= \mathbf{a}^T\Sigma\mathbf{a}
\end{aligned}
$$

(b) Because the variance of a random variable is always positive, we can assert that for any $\mathbf{a} \in \mathbb{R}^n, \mathbf{a} \neq 0$, $Var(\mathbf{a}^T\mathbf{x}) > 0$. But we just showed $Var(\mathbf{a}^T\mathbf{x}) = \mathbf{a}^T\Sigma\mathbf{a}$, hence for any $\mathbf{a} \in \mathbb{R}^n, \mathbf{a} \neq 0$, $\mathbf{a}^T\Sigma\mathbf{a} > 0$, so by definition Σ is positive definite.

Problems 1, Question 3: The ridge regression model in lecture 1 is an example of regularlization. For a dataset $(\mathbf{t}, \mathbf{X})$ consisting of targets $\mathbf{t} = (t_1, \ldots, t_N)$ and features $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)'$, we wish to build a model $y(\mathbf{x}_n, \mathbf{w})$ by minimizing

$$L(\mathbf{w}) = \frac{1}{2}(\mathbf{y} - \mathbf{t})'(\mathbf{y} - \mathbf{t}) + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

with respect to $\mathbf{w}$, where $\mathbf{y} = \mathbf{X}\mathbf{w}$ is the vector of predicted targets and $\mathbf{X}$ contains a columnn of ones.

(a) Show that when $\lambda = 0$ (no regularization) the solution to the above is the *least squares* weights defined in lecture (slide above the ridge regression slide). Use the results about gradients from the previous question to take the gradient of $L(\mathbf{w})$, set it to zero, and solve.

(b) Show that the solution for $\lambda \in [0, \infty)$ is given by

$$\hat{\mathbf{w}} = \left(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}\right)^{-1} \mathbf{X}'\mathbf{t}$$

where $\mathbf{I}$ is the identity matrix of appropriate size.

(c) *Optional*: Show that if $\mathbf{X}$ does not have full column rank, then the unique un-regularlized least squares solution does not exist.

(d) *Optional*: Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ have rank $q < p$. Show that the ridge regression solution you derived above *does* exist, i.e. show $rank\left(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}\right) = p$.

Solution:

(a) For $\lambda = 0$, the loss function is the same as the loss function for linear regression, so of course they have the same solution. You could also do b) and then plug in $\lambda = 0$ (as we'll do here).

(b) The gradient with respect to $\mathbf{w}$ is

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{X}'(\mathbf{X}\mathbf{w} - \mathbf{t}) + \lambda\mathbf{w}$$

Setting to zero and solving gives the system of equations

$$\left(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}\right)\mathbf{w} = \mathbf{X}'\mathbf{t}$$

which has unique solution

$$\hat{\mathbf{w}} = \left(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}\right)^{-1} \mathbf{X}'\mathbf{t}$$

(c) If $\mathbf{X}$ does not have full column rank, $rank(\mathbf{X}) < p$, then $rank(\mathbf{X}'\mathbf{X}) < p$ as well, and $\mathbf{X}'\mathbf{X}$ is not invertible. Hence the system of equations

$$\left(\mathbf{X}'\mathbf{X}\right)\mathbf{w} = \mathbf{X}'\mathbf{t}$$

has infinitely many solutions, i.e. $\hat{\mathbf{w}}$ is not unique.

(d) The matrix $\lambda\mathbf{I}$ is full rank. In general, intuitively, it shouldn't be possible to add one matrix to another and have a result be *lower* rank. We can prove this by computing the singular values of $(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})$ and showing they are all non-zero. Linear algebra such as this is very important, but outside the scope of this course.

Problems 2, Question 5: Consider the exponential family form as given in the last few lecture slides of lecture 2:

$$L(\eta|\mathbf{X}) = \left(\prod_{i=1}^{n} h(\mathbf{x})\right) g(\eta)^n \exp\left(\eta^T \sum_{i=1}^{n} u(\mathbf{x}_i)\right)$$

(a) Write down the log-likelihood, which is just the log of the above expression.

(b) Show that the univariate Gaussian distribution can be written in this form, and give explicit expressions for η, $u(x)$, $h(x)$ and $g(\eta)$ in terms of x, μ, σ. This question is done for you in the slides, but do it again yourself.

(c) Recall for the univariate Gaussian that $E(x) = \mu$ and $E(x^2) = \sigma^2 + \mu^2$. Verify this is true using the formula for the exponential family,

$$-\nabla \log g(\eta) = E(u(x))$$

Note how much easier that was than integrating the density function to find the moments.

(d) Now, repeat the above for the *multivariate* Gaussian: find the exponential family form, giving explicit expressions for the η, $u(x)$, $h(x)$ and $g(\eta)$ in terms of $\mathbf{x}, \boldsymbol{\mu}, \Sigma$.

(e) Find $E(\mathbf{x})$ and $E(\mathbf{x}\mathbf{x}^T)$. Again, use the exponential family identity from (c). After, write down the equivalent integral that you just solved, and pat yourself on the back.

(f) Find the maximum likelihood estimators for η, and then use that to find the maximum likelihood estimators for $\boldsymbol{\mu}$ and Σ. Use the identity at the very end of the lecture slides,

$$-\nabla \log g(\hat{\eta}) = \frac{1}{n}\sum_{i=1}^{n} u(\mathbf{x}_i)$$

to make your life way easier.

Solution:

(a)

$$\ell(\eta|\mathbf{X}) = \left(\sum_{i=1}^{n} \log h(\mathbf{x})\right) + n \log g(\eta) + \eta^T \sum_{i=1}^{n} u(\mathbf{x}_i)$$

(b) For the univariate Gaussian,

$$\begin{aligned} h(\mathbf{x}) &= -\frac{n}{2}\log 2\pi \\ \log g(\boldsymbol{\eta}) &= -\frac{1}{2}\log\sigma^2 + \frac{1}{2}\frac{\mu^2}{\sigma^2} \\ \boldsymbol{\eta} &= \begin{pmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{pmatrix} \\ u_i &= \begin{pmatrix} x_i \\ x_i^2 \end{pmatrix} \end{aligned}$$

(c) Rearrange to get $\boldsymbol{\eta}$ in terms of μ, σ^2:

$$\begin{aligned} \mu &= -\frac{\eta_1}{2\eta_2} \\ \sigma^2 &= -\frac{1}{2\eta_2} \end{aligned}$$

Take the gradient of $\log g(\boldsymbol{\eta})$ to obtain

$$\begin{aligned} \frac{\partial \log g(\boldsymbol{\eta})}{\partial \eta_1} &= -\frac{\eta_1}{2\eta_2} = \mu \\ \frac{\partial \log g(\boldsymbol{\eta})}{\partial \eta_2} &= -\frac{1}{2\eta_2} + \left(\frac{\eta_1}{2\eta_2}\right) = \sigma^2 + \mu^2 \end{aligned}$$

(d) For the multivariate Gaussian, the density is

$$f(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{|2\pi\Sigma|}} \exp\left(-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}\right)' \Sigma^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right)$$

which can be written as

$$\begin{aligned} f(\mathbf{x}|\boldsymbol{\mu}, \Sigma) &= h(\mathbf{x}) g(\boldsymbol{\eta}) \exp\left(\boldsymbol{\eta}'\mathbf{u}\right) \\ h(\mathbf{x}) &= (2\pi)^p \\ \log g(\boldsymbol{\eta}) &= -\frac{1}{4}\eta_1'\eta_2^{-1}\eta_1 - \frac{1}{2}\log\left|-2\eta_2^{-1}\right| = \frac{1}{2}\boldsymbol{\mu}'\Sigma\boldsymbol{\mu} + \frac{1}{2}|\Sigma| \\ \boldsymbol{\eta} &= \begin{pmatrix} \Sigma^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\Sigma^{-1} \end{pmatrix} \\ \mathbf{u} &= \begin{pmatrix} \mathbf{x} \\ \mathbf{x}\mathbf{x}' \end{pmatrix} \end{aligned}$$

after some extremely messy algebra. You should focus more on how to use the exponential family form, including all the relevant identities for expectations and maximum likelihood, rather than on doing the algebra necessary to get the density into the proper form.

(e) The gradient of $\log g(\boldsymbol{\eta})$ is (using identities for vector/matrix calculus; look these up on wikipedia if needed):

$$\begin{aligned} \frac{\partial g}{\partial \eta_1} &= -\frac{1}{2}\eta_2^{-1}\eta_1 = \boldsymbol{\mu} \\ \frac{\partial g}{\partial \eta_2} &= \frac{1}{4}\eta_2^{-1}\eta_1\eta_1'\eta_2^{-1} - \frac{1}{2}\eta_2^{-1} = \Sigma + \boldsymbol{\mu}\boldsymbol{\mu}' \end{aligned}$$

You should use the "trace-trick", $\mathbf{x}'\mathbf{A}\mathbf{x} = tr(\mathbf{A}\mathbf{x}\mathbf{x}')$, since a scalar is its own trace and the trace obeys cyclic permutations. That's a pretty popular trick in these types of questions.

(f) We just found the gradient of $\log g(\boldsymbol{\eta})$. Set it equal to the sample mean of the sufficient statistic and solve for η_1, η_2, to obtain the MLE for η_1, η_2. Rearrange these expressions to obtain

$$\begin{aligned} \hat{\boldsymbol{\mu}} &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i = \bar{\mathbf{x}} \\ \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}' + \hat{\Sigma} &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i \\ \implies \hat{\Sigma} &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i - \bar{\mathbf{x}}\bar{\mathbf{x}}' \end{aligned}$$

You already did all the equation solving and rearranging in the previous question; maximum likelihood estimation for the exponential family ends up just being the replacement of population moments with sample moments.

Problems 3, Question 3: Let

$$y \sim Binom(n, p)$$

where n is known and p is an unknown parameter. Put a prior on p,

$$p \sim Beta(\alpha, \beta)$$

where the Beta distribution has pdf

$$f(p|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

Find the posterior distribution of $p|y$. Explain why this means that the Beta distribution is the *conjugate prior* for the Binomial.

Solution: The prior is

$$f(p|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

The likelihood is

$$f(y|p) = \binom{n}{y} p^y (1-p)^{n-y}$$

The posterior is

$$f(p|y, \alpha, \beta) = \frac{f(y|p) \times f(p|\alpha, \beta)}{\int_0^1 f(y|p) \times f(p|\alpha, \beta) dp}$$

Because I told you that this prior is *conjugate* for the posterior, you know that the posterior takes the same functional form as the prior. This should give you confidence that the complicated-looking integral in the denominator is simple to evaluate. Indeed, note that because the beta distribution is, in fact, a distribution, we must have

$$\int_0^1 f(p|\alpha, \beta) dp = \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} dp = 1$$

Which means that

$$\int_0^1 p^{\alpha-1}(1-p)^{\beta-1} dp = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

The integral in the denominator is

$$\begin{aligned} &\int_0^1 \binom{n}{y} p^y (1-p)^{n-y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} dp \\ &= \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 p^{y+\alpha-1}(1-p)^{n-y+\beta-1} dp \\ &= \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\beta+\alpha)} \end{aligned}$$

The whole posterior therefore evaluates to

$$\begin{aligned} f(p|y, \alpha, \beta) &= \frac{\frac{\binom{n}{y}\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1}}{\frac{\binom{n}{y}\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\beta+\alpha)}} \\ &= \frac{\Gamma(n+\beta+\alpha)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1} \\ &\sim Beta(y+\alpha, n-y+\beta) \end{aligned}$$

This is what is meant by *conjugate prior*: given the data, the posterior is the same functional form as the prior, with parameters that depend on both the data and the hyperparameters of the prior.