Solutions to the STA414 / STA2104 Midterm Test

Dept of Statistical Sciences, University of Toronto
13 February 2017

1. **Cross-validation for Regression [10 Points]**

   With one point removed, the dataset will have only two points, so with λ close to zero, the regression line will pass through these two points, whereas with λ very large, the regression line will be horizontal, at the level equal to the mean of the two responses.

   For λ close to zero, we see that leaving out points from left to right gives a squared error of $3^2, 1.5^2$, and 3^2, for a total of 20.25. For λ very large, leaving out points from left to right gives squared errors of 0, 1.5^2, and 1.5^2, for a total of 4.5. So based on this cross-validation assessment, we would prefer the very large value of λ.

   (With thanks to Radford Neal for a similar question.)

2. **Bayesian Inference [15 Points]**

   (a) **[5 Points]**
   The μ^x factor represents the fact that x trials were successful, and each trial's probability of success is μ. It's the probability of multiple Bernoulli trials.
   Similarly, the $(1-\mu)^r$ factor represents the fact that r trials were unsuccessful, and each trial's probability of failure is $1-\mu$.
   The $\binom{x+r-1}{x}$ factor represents the number of ways of arranging the trials before the last one. The last trial, trial $x+r$, is always a failure. The $x+r-1$ preceding trials can have arbitrary order.

   (b) **[10 Points]** By Bayes' theorem,

   $$p(\mu|x) = \frac{p(x|\mu)p(\mu)}{p(x)}$$

   where

   $$\begin{aligned} p(x|\mu) &= \prod_{n=1}^{N} \binom{x_n + r - 1}{x_n} \mu^{x_n} (1-\mu)^r \\ p(\mu) &= \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \mu^{\alpha_0 - 1} (1-\mu)^{\beta_0 - 1} \\ p(x) &= \int_{\mu} p(x|\mu)p(\mu) d\mu. \end{aligned}$$

   Because of conjugacy, the above posterior will also take the form:

   $$p(\mu|x) \propto \mu^{\alpha_N - 1}(1-\mu)^{\beta_N - 1}. \tag{1}$$

We can compare (1) to its preceding equations. Collecting therein the corresponding factors with identical bases,

$$\mathrm{p}(\mu|x) \propto \mu^{\alpha_0 - 1 + \sum x_n}(1-\mu)^{Nr+\beta_0-1}.$$

Comparing this with (1) yields $\alpha_N = \alpha_0 + \sum x_n$ and $\beta_N = Nr + \beta_0$.

Interpretation:

An α term in the β distribution, such as the α_N term we derived, represents the shape parameter which pulls the posterior β distribution towards $\mu = 1$ (high probability of success). If many heads/successful data are collected in each draw, x_n tends to be large; the evaluation of our expression for α_N increases and therefore the distribution shifts towards $\mu = 1$.

In contrast, the β_N term represents the shape parameter which pulls the posterior β distribution towards $\mu = 0$ (high probability of failure). Note that the product Nr indicates the total number of failures seen across all draws.

Thirdly, from the expressions for both terms we see that as more draws are conducted (N becomes large), the priors α_0 and β_0 become dwarved.

3. **The Bernoulli distribution [15 Points]**

$$\begin{aligned}
\mathrm{KL}(p\,||\,q) &= \sum_{x_1,x_2} p(x_1,x_2)\,\ln\frac{p(x_1,x_2)}{q(x_1,x_2)} \\
&= \sum_{x_1,x_2} p(x_1,x_2)\,\ln\frac{p(x_1,x_2)}{q(x_1)q(x_2)} \\
&= \sum_{x_1,x_2} p(x_1,x_2)\,\ln p(x_1,x_2) - \sum_{x_1,x_2} p(x_1,x_2)\,\ln q(x_1)q(x_2) \\
&= c - \sum_{x_1,x_2} p(x_1,x_2)\,\ln q(x_1)q(x_2) \\
&= c - \sum_{x_1,x_2} p(x_1,x_2)\,\ln q(x_1) - \sum_{x_1,x_2} p(x_1,x_2)\,\ln q(x_2) \\
&= c - \sum_{x_1} p(x_1)\,\ln q(x_1) - \sum_{x_2} p(x_2)\,\ln q(x_2)
\end{aligned}$$

Now the optimization is separate for $q(x_1)$ and $q(x_2)$.

Need to solve:

$$\begin{aligned}
&\arg\max_{\mu 1} \sum_{x_1} p(x_1)\,\ln q(x_1) \\
&= \arg\max_{\mu 1} \left[p(x_1=0)\ln q(x_1=0) + p(x_1=0)\ln q(x_1=0)\right]
\end{aligned}$$

Setting the derivative to zero gives $q(x_1=0) = p(x_1=1)$, which is the same as $\mu_1 = E_p[x_1]$.

4. **Manipulating Gaussians [10 Points]**

Let's list some easy properties:

$$
\begin{aligned}
E[x_0] &= 0 \\
E[x_1] &= aE[x_0] = 0 \\
E[x_2] &= bE[x_0] = 0 \\
V[x_0] &= \sigma^2 \\
V[x_1] &= a^2 V[x_0] + \sigma^2 = (1 + a^2)\sigma^2 \\
V[x_2] &= b^2 V[x_0] + \sigma^2 = (1 + b^2)\sigma^2 \\
E[x_1 | x_0] &= a x_0 \\
E[x_2 | x_0] &= b x_0
\end{aligned}
$$

Now consider the covariance:

$$
\begin{aligned}
\mathrm{Cov}[x_1, x_2] &= E[(x_1 - E[x_1])(x_2 - E[x_2])] \\
&= E[x_1 x_2] \\
&= \int\int\int x_1 x_2 p(x_0, x_1, x_2) dx_0 dx_1 dx_2 \\
&= \int\int\int x_1 x_2 p(x_1|x_0) p(x_2|x_0) p(x_0) dx_0 dx_1 dx_2 \\
&= \int\int x_1 x_2 p(x_1|x_0) \int p(x_2|x_0) p(x_0) dx_2 dx_0 dx_1 \\
&= \int E[x_1|x_0] E[x_2|x_0] p(x_0) dx_0 \\
&= \int a x_0 b x_0 p(x_0) dx_0 \\
&= abV[x_0] \\
&= ab\sigma^2
\end{aligned}
$$

So x_1, x_2 have mean $[0, 0]$, variance $[(1 + a^2)\sigma^2, (1 + b^2)\sigma^2]$, and covariance $ab\sigma^2$.

They are also jointly normally distributed. Recall from lecture 2 that the marginal distributions of a multivariate normal distribution are themselves normal. In reverse, the joint pdf of a set of normal distributions is not always a multivariate normal distribution; it often is, as in the present case, but a proof wasn't necessary here.

5. **Linear Binary Classification Models [13 Points]**

(a) **[3 Points]** Letting $f(\cdot)$ be an arbitrary, possibly nonlinear, activation function,

$$\mathrm{p}(c = 1|\mathbf{x}, \mathbf{w}) = f(\mathbf{x}^T\mathbf{w} + w_0)$$

and naturally $\mathrm{p}(c = 0|\mathbf{x}, \mathbf{w}) = 1 - \mathrm{p}(c = 1|\mathbf{x}, \mathbf{w})$.

(b) **[3 Points]** The data are linearly separable. If $f(\cdot)$ yields a sharp decision boundary (high slope in between the two classes) then $\mathrm{p}(t_i|\mathbf{x}_i, \mathbf{w}) \to 1$ for each datum. The product is $(\sim 1)^3 \approx 1$.

(c) **[3 Points]** The straight-line decision boundary gives either 1, 2, or 3 misclassifications. For the case of 1 misclassification, the maximum-likelihood boundary will be a 45-degree line, with likelihood $\mathcal{L} \approx (0.5 - 3a)(0.5 + a)^3$ for some nudge a in the logistic regression's output. For the case of 2 misclassifications, the ML-boundary is a vertical or horizontal line at $x_1 = 0.5$ or $x_2 = 0.5$ respectively, with $\mathcal{L} = (0.5 - a)^2(0.5 + a)^2$. For all the above, optimizing yields $a = 0$ and $\mathcal{L} = 1/16$.

(d) **[4 Points]** Possible answers:

- $\boldsymbol{\phi}_1(\mathbf{x}) = \exp\left[-\frac{(\mathbf{x}-(0,0)^T)^2}{2(0.2)^2}\right]$ and $\boldsymbol{\phi}_2(\mathbf{x}) = \exp\left[-\frac{(\mathbf{x}-(1,1)^T)^2}{2(0.2)^2}\right]$
- $\boldsymbol{\phi}_1(\mathbf{x}) = (x_1 - 0.5)(x_2 - 0.5)$
- $\boldsymbol{\phi}_1(\mathbf{x}) = x_1$, $\boldsymbol{\phi}_2(\mathbf{x}) = x_2$, $\boldsymbol{\phi}_3(\mathbf{x}) = x_1x_2$, and $\boldsymbol{\phi}_4(\mathbf{x}) = 1$
- Many others