MLPR Tutorial Sheet 6

Reminder: If you need more guidance to get started on a question, seek clarifications and hints on the class forum. Move on if you're getting stuck on a part for a long time. Full answers will be released at the end of Thursday 2016-11-10.

---

1. **More practice with Gaussians:**

   N noisy independent observations are made of an unknown scalar quantity m:

   $$x^{(n)} \sim \mathcal{N}(m, \sigma^2).$$

   a) I don't give you the data, but tell you the mean of the observations:

   $$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x^{(n)}.$$

   What is the likelihood of m given only this mean $\bar{x}$? That is, what is $p(\bar{x}\,|\,m)$?

   b) A *sufficient statistic* is a summary of some data that contains all of the information about a parameter.

   i) Show whether $\bar{x}$ is a sufficient statistic of the observations, assuming we know the noise variance σ^2. That is, show whether $p(m\,|\,\bar{x}) = p(m\,|\,\{x^{(n)}\}_{n=1}^{N})$.

   ii) If we don't know the noise variance σ^2, show whether $\bar{x}$ is a sufficient statistic to estimate it.

2. **Conjugate priors:**

   A *conjugate prior* for a likelihood function is a prior where the posterior is a distribution in the same family as the prior. For example, a Gaussian prior on the mean of a Gaussian distribution is conjugate to Gaussian observations of that mean.

   a) The *inverse-gamma distribution* is a distribution over positive numbers. It's often used to put a prior on the variance of a Gaussian distribution, because it's a conjugate prior.

   The inverse-gamma distribution has pdf (as cribbed from Wikipedia):

   $$p(z\,|\,\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{-\alpha-1} \exp\left(-\frac{\beta}{z}\right), \qquad \text{with } \alpha > 0,\ \ \beta > 0,$$

   where $\Gamma(\cdot)$ is a gamma function. (Numerical libraries often come with a routine to evaluate the log of the gamma function.)

   Assume we obtain N observations from a zero-mean Gaussian with unknown variance,

   $$x^{(n)} \sim \mathcal{N}(0, \sigma^2), \quad n = 1 \ldots N,$$

   and that we place an inverse-gamma prior with parameters α and β on the variance. Show that the posterior over the variance is inverse-gamma, and find its parameters.

   Hint: you can assume that the posterior distribution is a distribution; it normalizes to one. You don't need to keep track of normalization constants, or do any integration. Simply show that the posterior matches the functional form of the inverse-gamma, and then you know the normalization (if you need it) by comparison to the pdf given.

b) i) If a conjugate prior exists, then the data can be replaced with sufficient statistics. Can you explain why?

ii) Explain whether there could be a conjugate prior for the hard classifier:

$$P(y=1\,|\,\mathbf{x},\mathbf{w}) = \Theta(\mathbf{w}^\top\mathbf{x}+b) = \begin{cases} 1 & \mathbf{w}^\top\mathbf{x}+b>0 \\ 0 & \text{otherwise.} \end{cases}$$

This question is intended as a tutorial discussion point. It might be hard to write down a mathematically rigorous argument. But can you explain whether it is easy to represent beliefs about the weights of a classifier in a fixed-size statistic, regardless of what data you gather? A picture may help.

3. **Regression with input-dependent noise:**

In lectures we turned a model of a function into a probabilistic model of real-valued outputs by modelling the residuals as Gaussian noise:

$$p(y\,|\,\mathbf{x},\theta) = \mathcal{N}(y;\, f(\mathbf{x};\theta),\, \sigma^2).$$

The noise variance σ^2 is often assumed to be a constant, but it could be a function of the input location $\mathbf{x}$. (A "heteroscedastic" model.)

A flexible model could set the variance using a neural network:

$$\sigma(\mathbf{x})^2 = \exp(W^{(\sigma)}\mathbf{h}(\mathbf{x};\theta) + b^{(\sigma)}),$$

where $\mathbf{h}$ is a vector of hidden unit values. These could be hidden units from the neural network used to compute function $f(\mathbf{x};\theta)$, or there could be a separate network to model the variances.

a) Assume that $\mathbf{h}$ is the final layer of the same neural network used to compute f. How would we modify the training procedure for a neural network that fits f by least squares, to fit this new model?

b) In the suggestion above, the activation $a^{(\sigma)} = W^{(\sigma)}\mathbf{h} + b^{(\sigma)}$ sets the log of the variance of the observations.

i) Why not set the variance directly to this activation value, $\sigma^2 = a$?

ii) Harder (I don't know if you'll have an answer, but I'm curious to find out): Why not set the variance to the square of this activation value, $\sigma^2 = a^2$?

c) Given a test input $\mathbf{x}^{(*)}$, the model above outputs both a guess of an output, $f(\mathbf{x}^{(*)})$, and an 'error bar' $\sigma(\mathbf{x}^{(*)})$, which indicates how wrong the guess could be.

The Bayesian linear regression and Gaussian process models covered in lectures also give error bars on their predictions. What are the pros and cons of the neural network approach in this question? Would you use this neural network to help guide experiment design?