CS 645 Problem Set One, Selected Solutions and Comments

1 (a)

There are a variety of features and components that form the Bayesian approach, with different authors emphasizing different sets of them. When contrasted with the frequentist approach, a key part that is often emphasized is general quantification of uncertainty which allows for broader application such as a calculus of uncertainty that supports subjective ideas of events that are rare or have not yet been observed. (This point was noted by most who attempted this question).

More technical components include:

i) Parameters are considered random variables
ii) Prior probabilities are exploited, either to incorporate specific prior information, do sequential updates, or as a simple way to regularize (smooth).
iii) The posterior distribution is computed using Bayes' rule.
iv) The entire posterior is used to estimate quantities such as risk, cost, or class. (A true Bayesian tends to defer a point estimate as long as they can).

1 (b)

Informally, a generative model tells a story about how the data was generated. More formally, it specifies the joint probability of the relevant parts of the world. Typically it is a clearly parametric model. From the perspective of learning what generative models are (and they form a key part of this course) one should consider how, in specific instances, to "tell" the generative story at various levels. At the least descriptive level, i.i.d. data is generated by repeated samples from joint distribution, but invariably there is more structure. For example, for the points forming data arround a line, the story could be as follows. First, a sample is taken from the prior distribution over parameters to create an instance of the parameters. Then a sample is taken from the prior distribution over X. Then the value of Y is computed using the parameters to compute the mean. Then a Gaussian is sampled with that mean and the variance (where the variance comes from would be described in a specific case. Note that to tell a generative story of line data, we had to have a prior over X. Often X is simply thought of as a "input" variable, but then the model is not truly generative.

1 (c)

I may have created a bit of confusion in this problem because I think of the discriminative *approach* as including both discriminative modeling (described in labeled paragraph (b) on page 43) and using a discriminative function (described in labeled paragraph (c) on page 43).

A discriminative model directly models the conditional probability of something of interest (e.g. class) based on the data. The origin of the data point, or how likely it is not

of interest. Invariably this is in the context of decision theory where what is of interest is (not surprisingly) the decision. When these are combined, this approach resembles creating a simple discriminative function which directly maps the data to the decision. Probabilities may or may not have been involved explicitly or implicitly in developing the function (this will become more clear as the course progresses and we see examples).

Further comments made in class:

A good generative model supports all tasks, but finding it may be difficult, and wasteful in the sense that some of the modeling power of the data is expended on features of the joint distribution which are not useful for the decision task at hand. This is most easily explained in the case of a discriminative function which can draw the boundary based on data, without understanding anything about the probability distribution of the data. It was pointed out that one weakness of the discriminative approach is that it cannot identify outliers. Also, it cannot be used to generate data (e.g. for simulations).

2.

The key thing realized by most who attempted this question is that the volume of the shell of width epsilon exponentially with D, but the density over this volume is stable. Hence the probability mass as a function of radius gets pushed towards the edge of the Gaussian hyper "ball".

3.

The main point that I was "fishing for" is that the estimate for the new value is not based on a point estimate for the parameters, but based on all possible values, weighted by how likely they are. This kind of equation often is the meat of the application of the Bayesian method, where we prefer to work with posterior distributions for as long as is practical, before committing to a point estimate.

4.

I saw no real problems with this question. Most students attempted it, and solutions were generally good. The step that needed to be filled involved expanding the terms, noting that the expectations that are already available are constants in the integration (or sum) that comes from the outer expectation, and because expectation is an integral (sum), it is linear. The following is one version of the details provided by one student:

$$
\begin{aligned}
cov[x, y] &= \mathbb{E}_{x,y}[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}] \\
&= \mathbb{E}_{x,y}\{xy - x\mathbb{E}[y] - y\mathbb{E}[x] + \mathbb{E}[x]\mathbb{E}[y]\} \\
&= \mathbb{E}_{x,y}\{xy\} - \mathbb{E}_{x,y}\{x\mathbb{E}[y]\} - \mathbb{E}_{x,y}\{y\mathbb{E}[x]\} + \mathbb{E}_{x,y}\{\mathbb{E}[x]\mathbb{E}[y]\} \\
&= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] - \mathbb{E}[y]\mathbb{E}[x] + \mathbb{E}[x]\mathbb{E}[y] \\
&= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]
\end{aligned}
$$

For the problem itself, the key algebra of course comes from the definition of independent that specifies that p(x,y)=p(x)p(y). One way to handle the details is:

Assume x and y are independent, then

$$
\begin{aligned}
\mathbb{E}_{x,y}[xy] &= \int_Y \int_X x\, y\, p(x, y)\, dx\, dy \\
&= \int_Y y\, p(y) \int_X x\, p(x)\, dx\, dy \\
&= \mathbb{E}[x] \int_Y y\, p(y)\, dy \\
&= \mathbb{E}[x]\mathbb{E}[y]
\end{aligned}
$$

5.

We covered this one in detail in class.

6.

We also covered this on in detail in class. To re-iterate, the point of this question is that least squares error is tightly linked with maximum likelihood estimation of the parameters, itself being achieved by applying Bayes rule.

7.

Not many attempted this question, which is fortunate, because I actually had meant to write “problem 1.27” (sorry!). I have corrected this for future incarnations of this course. (The few that did 1.23 will get credit under the substitution rule).