Theory of Inference: Homework 2

1. Study the statement and proof of the Bayes Rule theorem. Now state and prove the theorem in the special case where the choice of action does not affect the client's beliefs about X.

   Answer. Define

   $$\delta^*(y) := \operatorname*{argmin}_{a \in \mathcal{A}} \mathrm{E}\{L(a, X) \mid Y \doteq y\},$$

   note that there is no '$; a$' in the expectation, because the action does not affect beliefs about X. We want to show that, for arbitrary rule δ,

   $$\mathrm{E}\{L(\delta(Y), X)\} \geq \mathrm{E}\{L(\delta^*(Y), X)\},$$

   which would prove that δ^* was a Bayes Rule. Thus:

   $$\begin{aligned}
   \mathrm{E}\{L(\delta(Y), X)\} &= \sum\nolimits_x \sum\nolimits_y L(\delta(y), x)\, \mathrm{p}(x, y) && \text{FTP} \\
   &= \sum\nolimits_y \sum\nolimits_x L(\delta(y), x)\, \mathrm{p}(x \mid y)\, \mathrm{p}(y) && \text{definition of } \mathrm{p}(x \mid y) \\
   &\geq \sum\nolimits_y \left\{ \min_{a \in \mathcal{A}} \sum\nolimits_x L(a, x)\, \mathrm{p}(x \mid y) \right\} \mathrm{p}(y) \\
   &= \sum\nolimits_y \left\{ \min_{a \in \mathcal{A}} \mathrm{E}\{L(a, X) \mid Y \doteq y\} \right\} \mathrm{p}(y) && \text{CFTP} \\
   &= \sum\nolimits_y \mathrm{E}\{L(\delta^*(y), X) \mid Y \doteq y\}\, \mathrm{p}(y) && \text{by definition} \\
   &= \sum\nolimits_y \left\{ \sum\nolimits_x L(\delta^*(y), x)\, \mathrm{p}(x \mid y) \right\} \mathrm{p}(y) && \text{CFTP again} \\
   &= \sum\nolimits_y \sum\nolimits_x L(\delta^*(y), x)\, \mathrm{p}(x, y) && \mathrm{p}(x \mid y) \text{ again} \\
   &= \mathrm{E}\{L(\delta^*(Y), X)\} && \text{FTP again}
   \end{aligned}$$

   as required. This is overkill on the number of steps; about four would do.

2. A volcano can be either inactive or active; it is active with probability θ. If it is inactive its eruption rate is zero. If it is active, its eruption rate

has a Gamma distribution with shape parameter α and rate parameter β. Let λ be its eruption rate. Show that

$$\Pr(\lambda \dot{\leq} v) = 1 - \theta + F(v; \alpha, \beta) \cdot \theta$$

for $v \geq 0$, and zero otherwise, where F is the distribution function of the Gamma distribution. Hint: introduce the random quantity $A \in \{0, 1\}$, where $A = 1$ exactly when the volcano is active, and use the Law of Total Probability.

Answer. Introduce A as in the hint, so that $\{A \doteq 0, A \doteq 1\}$ is a belief partition. Let $\gamma \sim \text{Gamma}(\cdot\,; \alpha, \beta)$. Then $\lambda = A\gamma$, and

$$\begin{aligned}
\Pr(\lambda \dot{\leq} v) &= \Pr(\lambda \dot{\leq} v \mid A \doteq 0) \Pr(A \doteq 0) + \Pr(\lambda \dot{\leq} v \mid A \doteq 1) \Pr(A \doteq 1) && \text{LTP} \\
&= \Pr(A\gamma \dot{\leq} v \mid A \doteq 0) \Pr(A \doteq 0) + \Pr(A\gamma \dot{\leq} v \mid A \doteq 1) \Pr(A \doteq 1) \\
&= \Pr(0 \dot{\leq} v \mid A \doteq 0) \Pr(A \doteq 0) + \Pr(\gamma \dot{\leq} v \mid A \doteq 1) \Pr(A \doteq 1) && \text{by TOWIK} \\
&= 1 \cdot \Pr(A \doteq 0) + F(v; \alpha, \beta) \cdot \Pr(A \doteq 1) && \text{as } v \geq 0 \\
&= 1 - \theta + F(v; \alpha, \beta) \cdot \theta
\end{aligned}$$

as required.

3. Here is an exam-style revision question on the first strand of lectures.

   (a) The Fundamental Theorem of Prevision (FTP) is an if-and-only-if theorem. Pick one branch (either if or only-if), state it, and prove it. [5 marks]

   Answer. All of these answers are covered in the notes; I'll produce off-the-cuff answers here without reference to the notes—just for fun.

   One branch of the FTP states that if E is coherent then there exists a $\boldsymbol{p} \in \mathbb{S}^{s-1}$ for which $\mathrm{E}\{g(\boldsymbol{X})\} = \sum_j g(\boldsymbol{x}^{(j)}) \cdot p_j$ for any specified g, where s is the dimension of the realm of $\boldsymbol{X}$, which is $\boldsymbol{\mathcal{X}} := \{\boldsymbol{x}^{(j)}\}_{j=1}^{s}$. To prove this, note that

$$1 = \sum\nolimits_j \mathbb{1}_{\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}},$$

because $\boldsymbol{X}$ must take exactly one value from its realm. And then

$$\begin{aligned}
\mathrm{E}\{g(\boldsymbol{X})\} &= \mathrm{E}\{g(\boldsymbol{X}) \cdot 1\} \\
&= \mathrm{E}\left\{ g(\boldsymbol{X}) \sum\nolimits_j \mathbb{1}_{\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}} \right\} \\
&= \mathrm{E}\left\{ \sum\nolimits_j g(\boldsymbol{X}) \mathbb{1}_{\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}} \right\} \\
&= \mathrm{E}\left\{ \sum\nolimits_j g(\boldsymbol{x}^{(j)}) \mathbb{1}_{\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}} \right\} \quad \text{(make sure you understand this step)} \\
&= \sum\nolimits_j g(\boldsymbol{x}^{(j)}) \, \mathrm{E}\{\mathbb{1}_{\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}}\} \quad \text{by linearity} \\
&= \sum\nolimits_j g(\boldsymbol{x}^{(j)}) \cdot p_j
\end{aligned}$$

say. We have $p_j \geq 0$ by lower-boundedness, and also

$$1 = \mathrm{E}\{1\} = \mathrm{E}\left\{ \sum\nolimits_j \mathbb{1}_{\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}} \right\} = \sum\nolimits_j \mathrm{E}\{\mathbb{1}_{\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}}\} = \sum\nolimits_j p_j$$

by lower-boundedness and linearity, showing that $\boldsymbol{p} := (p_1, \ldots, p_s)$ is a point in $\mathbb{S}^{s-1}$, as required.

(b) Outline a model for data, distinguishing between random quantities, observables, and observations. Give an example of how an observable differs from the random quantity it measures, owing to limitations in the instrument. [5 marks]

Answer. Suppose we have m random quantities $\boldsymbol{X} := (X_1, \ldots, X_m)$. An 'observable' Y_i is a specified function of $\boldsymbol{X}$, say $g_i(\boldsymbol{X})$, whose value will at some point become known; its value is the 'observation' y_i^{obs}. 'Data' can collectively be represented as the truth of the random proposition

$$Q := q(\boldsymbol{X}) \quad \text{where} \quad q(\boldsymbol{x}) := \bigwedge_{i=1}^{n} \big(g_i(\boldsymbol{x}) \doteq y_i^{\mathrm{obs}} \big).$$

In the simplest possible case, $g_i(\boldsymbol{x}) \leftarrow x_i$, in which case the data comprise observations on the first n random quantities: call this the *Simple Observation Model (SOM)*. More generally, Y_i might be a more complicated function of X_i, reflecting in part the measuring instrument's limitations. For example, a meter which only reads up to the value v has $g_i(\boldsymbol{x}) = \min\{x_i, v\}$.

(c) State and prove the Muddy Table theorem. Illustrate it with a diagram. [5 marks]

Answer. Let $q(\boldsymbol{x})$ be a first order sentence, and define the random proposition $Q := q(\boldsymbol{X})$. The Muddle Table theorem states that

$$\Pr(\boldsymbol{X} \doteq \boldsymbol{x} \mid Q) = \frac{\mathbb{1}_{q(\boldsymbol{x})}\, \mathrm{p}(\boldsymbol{x})}{\Pr(Q)} \quad \text{provided that } \Pr(Q) > 0,$$

where $\mathrm{p}(\cdot)$ represents the PMF of $\boldsymbol{X}$ and $\Pr(Q) = \sum_{\boldsymbol{x}} \mathbb{1}_{q(\boldsymbol{x})}\, \mathrm{p}(\boldsymbol{x})$ by the Fundamental Theorem of Prevision (FTP).

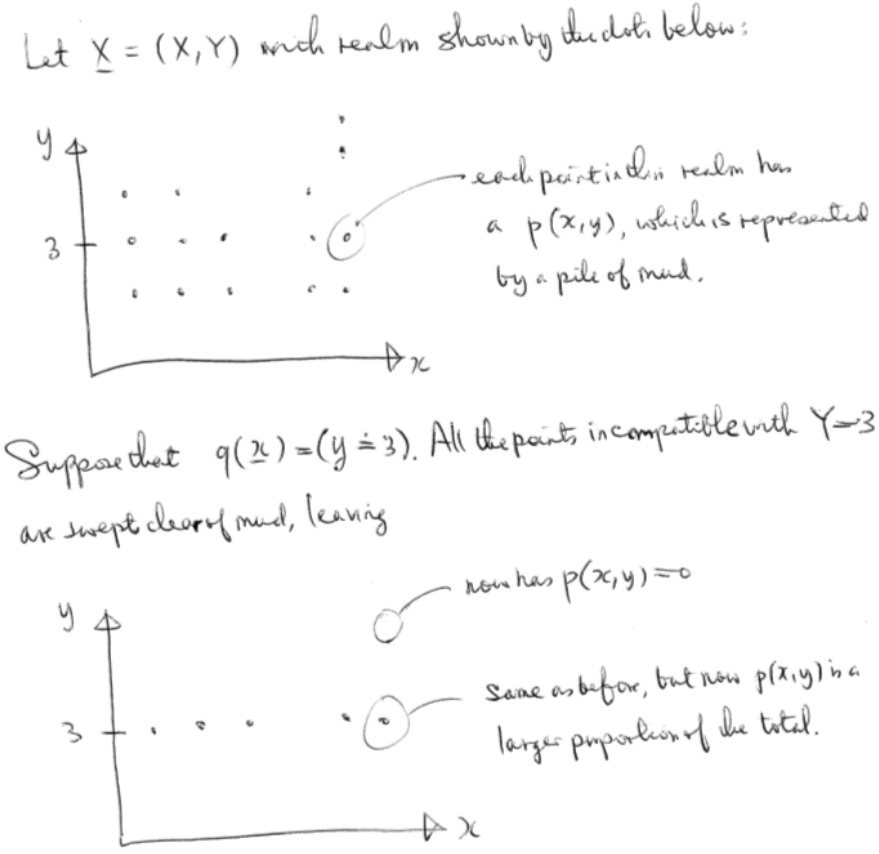
For the proof, take $\Pr(Q) > 0$ as stated, and then

$$\Pr(\boldsymbol{X} \doteq \boldsymbol{x} \mid Q) = \frac{\Pr(\boldsymbol{X} \doteq \boldsymbol{x}, Q)}{\Pr(Q)}$$

according to the definition of hypothetical/conditional probability. Evaluate the numerator using the FTP to get

$$\Pr(\boldsymbol{X} \doteq \boldsymbol{x}, Q) = \sum_{\boldsymbol{x}'} \mathbb{1}_{\boldsymbol{x}' \doteq \boldsymbol{x} \wedge q(\boldsymbol{x}')}\, \mathrm{p}(\boldsymbol{x}') = \mathbb{1}_{q(\boldsymbol{x})}\, \mathrm{p}(\boldsymbol{x})$$

as required. Here is my diagram:



(d) 'Bayesian conditionalisation' is the name given to a model for learning in which new information is incorporated into beliefs through conditioning. What are the attractive features of this model? Why does it *not* describe the typical practice of statistical inference? [10 marks]

Answer. Here are some bullet points for this answer:

- Bayesian conditionalisation asserts that beliefs are represented as probability statements and updated by conditioning. A set of propositions Ψ is already known to be true, and another proposition Q is newly learnt to be true, and is to be added to Ψ. Before knowing Q, beliefs have the form $\text{Bel}_\Psi(\cdot) = \Pr(\cdot \mid \Psi)$. After knowing Q, updated beliefs have the form $\text{Bel}_{Q \wedge \Psi}(\cdot) = \Pr(\cdot \mid Q, \Psi)$,

provided that $\Pr(Q \mid \Psi) > 0$.

- One attractive feature of this approach is that it has a very simple form: it is easy to compute the updated beliefs by applying sequential Bayes's theorem. Other attractive features include:
  i. It is a theorem that conditional probabilities represented as $\Pr(Q \mid R)$ in the defining relationship

  $$\Pr(Q, R) = \Pr(Q \mid R) \Pr(R)$$

  are indeed probabilities (i.e. satisfy the axioms of expectation), so this type of update is 'coherence preserving'.
  ii. As a specical case of the above, the update satisfies the minimal property of being logically consistent, so that if Q implies R then $\Pr(R \mid Q) = 1$.
  iii. The update is order-invariant, so that if $Q = Q_1 \wedge Q_2$, then the same beliefs arise from conditioning first on Q_1 and then on Q_2, or the other way around: in other words, only Ψ matters, not the order in which its elements were acquired.
- The learning model has a simple sequential form in which we start with some beliefs Ψ, acquire new knowledge Q, and then update to $Q \wedge \Psi$. Practically speaking, it can only work if we anticipate Q, in order that $\Pr(Q \mid \Psi) > 0$. It is very hard to anticipate everything that one might learn.
  More generally, in a statistical inference the nature of Q is itself shaped through dialogue with the client, and analysis of the datapool. Often the Q that is selected is only a small subset of the datapool, reflecting the statistician's and the client's limitations in specifying defensible beliefs about many of the elements of the datapool. This 'backwards and forwards' activity to choose Q is quite different from the simple sequential form represented in Bayesian conditionalisation.

Please hand in your answers for marking next Tue (3 Mar), at the lecture or by 5pm in the box outside my office door. I will return them in Thu's

lecture.

Jonathan Rougier
Feb 2015