Theory of Inference: Homework 3

1. Important but a bit tedious. Let $\boldsymbol{X} := \{X_1, \dots, X_m\} \in \boldsymbol{\mathcal{X}}$, as usual. Show that the PMF of the observables is

$$\mathrm{p}(\boldsymbol{y}) = \prod_{i=1}^{n} \mathrm{p}_X(y_i)$$

under the following conditions:

(a) Simple observational model (SOM), i.e.

$$Y_i = X_i \qquad i = 1, \dots, n,$$

(b) $X_1, \dots, X_m \overset{\text{iid}}{\sim} \mathrm{p}_X(\cdot)$, i.e.

$$\mathrm{p}(\boldsymbol{x}) = \prod_{i=1}^{m} \mathrm{p}_X(x_i).$$

Answer. Let's do this with $m \leftarrow 4$ and $n \leftarrow 2$.

$$\begin{aligned}
\mathrm{p}(y_1, y_2) &= \mathrm{E}\{\mathbb{1}_{Y_1 \doteq y_1 \wedge Y_2 \doteq y_2}\} && \text{Def. of prob.} \\
&= \mathrm{E}\{\mathbb{1}_{X_1 \doteq y_1 \wedge X_2 \doteq y_2}\} && \text{Cond. (a)} \\
&= \sum_{x_1} \cdots \sum_{x_4} \mathbb{1}_{x_1 \doteq y_1 \wedge x_2 \doteq y_2} \cdot \mathrm{p}(x_1, x_2, x_3, x_4) && \text{FTP} \\
&= \sum_{x_1} \cdots \sum_{x_4} \mathbb{1}_{x_1 \doteq y_1} \cdot \mathbb{1}_{x_2 \doteq y_2} \cdot \mathrm{p}_X(x_1) \cdots \mathrm{p}_X(x_4) && \text{Cond. (b)} \\
&= \sum_{x_1} \mathbb{1}_{x_1 \doteq y_1} \, \mathrm{p}_X(x_1) \cdot \sum_{x_2} \mathbb{1}_{x_2 \doteq y_2} \, \mathrm{p}_X(x_2) \cdot \sum_{x_3} \mathrm{p}_X(x_3) \cdot \sum_{x_4} \mathrm{p}_X(x_4) \\
&= \mathrm{p}_X(y_1) \cdot \mathrm{p}_X(y_2) \cdot 1 \cdot 1
\end{aligned}$$

as required.

2. Here is a exam-style revision question on decision theory and prediction. Each part is worth five marks.

(a) Describe the statistical framework for analysing decision problems, including decision rules. Define what it means for an action or a decision rule to be optimal.

Answer. As previously, these answers will be free-style, although they should not be too different from what you will find in the handouts and lecture notes.

Given a set of random quantities $\boldsymbol{X} \in \boldsymbol{\mathcal{X}}$, the two essential features of a decision problem are an action space $\mathcal{A} := \{a_1, \dots, a_k\}$ and a loss function $L : \mathcal{A} \times \boldsymbol{\mathcal{X}} \to \mathbb{R}$, where $L(a, \boldsymbol{x})$ is the loss incurred for choosing action a when $\boldsymbol{X}$ turns out to be $\boldsymbol{x}$. An action is optimal when it minimises the expected loss, i.e.

$$a^* = \operatorname*{argmin}_{a \in \mathcal{A}} \mathrm{E}\{L(a, \boldsymbol{X})\}.$$

In a more complicated situation, observables $\boldsymbol{Y} \in \boldsymbol{\mathcal{Y}}$ will be available before the action is chosen. In that case the task is to choose a decision rule $\delta : \boldsymbol{\mathcal{Y}} \to \mathcal{A}$, where $\delta(\boldsymbol{y})$ is the action selected if the observations are $\boldsymbol{y}$. A decision rule δ^* is optimal if it satisfies

$$\delta^* = \operatorname*{argmin}_{\delta \in \mathcal{D}} \mathrm{E}\{L[\delta(\boldsymbol{Y}), \boldsymbol{X}]\}$$

where $\mathcal{D}$ is the set of all possible decision rules.

(b) Consider the special case of a prediction problem: how does this differ from more general decision problems? Provide illustrations of prediction problems for the same random quantity X, but which are likely to differ in the prediction that is made.

Answer. In a prediction problem, the objective is to provide a point value for a random quantity, X say, either directly, or as a function of some observables, $\boldsymbol{y}$. The action space is $\mathcal{A} = \mathbb{R}$ (or possibly some subset of it), and the loss function is then $L : \mathbb{R} \times \mathcal{X} \to \mathbb{R}$. So the difference is mainly in the action space.

If X was the weight of a donkey assessed for the purposes of treatment, then, as discussed in the lectures, the loss from a wrong prediction can be asymmetric. Some drugs, such as wormers, incur a larger loss from an under-prediction than from an over-prediction; others, such as anaesthetics, incur a larger loss from over-prediction; the loss in both cases being measured in terms of the animal's health. So predictions made to minimise expected loss will be different in the two cases, with the predictions for wormers being systematically higher than the predictions for anaesthetics.

(c) Explain the motivation for producing 'generic' predictions, and justify the use of convex loss functions for such predictions.

Answer. In many decision problems, the increase in expected losses from making a slightly sub-optimal decision are outweighted by the increase in cost from hiring a statistician and doing a careful decision analysis. Also, of course, many people want to make a quick decision and are unaware that there is a framework within which they could take account of their uncertainties when choosing between actions. In these cases the simple rule is to replace expected loss over X with the loss evaluated at some predicted value for X, i.e. to hope that

$$\operatorname*{argmin}_{a \in \mathcal{A}} \mathrm{E}\{L(a, X)\} \approx \operatorname*{argmin}_{a \in \mathcal{A}} L(a, \hat{x})$$

where $\hat{x}$ is a prediction for X; there is a straightforward generalisation to many X's. In this case we need a 'generic' prediction for X which will work across many different decision problems.

When deciding on a loss function for a generic prediction, one natural feature, which would be widely accepted, is that small prediction errors are tolerable, but large ones are intolerable. This is captured by a convex loss function

$$L(x', x) = h(x' - x)$$

for some non-negative and convex h satisfying $h(0) = 0$. This loss function attributes zero loss to a correct prediction, a (relatively) small loss to a nearly-correct prediction, and a (relatively) large loss to a large mis-prediction. [A picture of h would be helpful here.]

(d) Let $\boldsymbol{Y}$ be a set of observables with the statistical model $\boldsymbol{Y} \sim \mathrm{p}(\cdot\,; \theta)$ for some $\theta \in \Omega$. Define what is meant by an 'estimator', and what it means for an estimator to be 'admissible' (for simplicity, assume that Ω is finite).

Answer. An estimator is a prediction rule for θ based on $\boldsymbol{y}$. Let Ω be a convex subset of $\mathbb{R}^p$ and let $\hat{\theta} : \boldsymbol{\mathcal{Y}} \to \mathbb{R}^p$ be any estimator. Define the risk function for a specified loss function $L : \mathbb{R}^p \times \Omega \to \mathbb{R}$ as

$$R(\delta, \theta) := \mathrm{E}\{L[\hat{\theta}(\boldsymbol{Y}), \theta]; \theta\}.$$

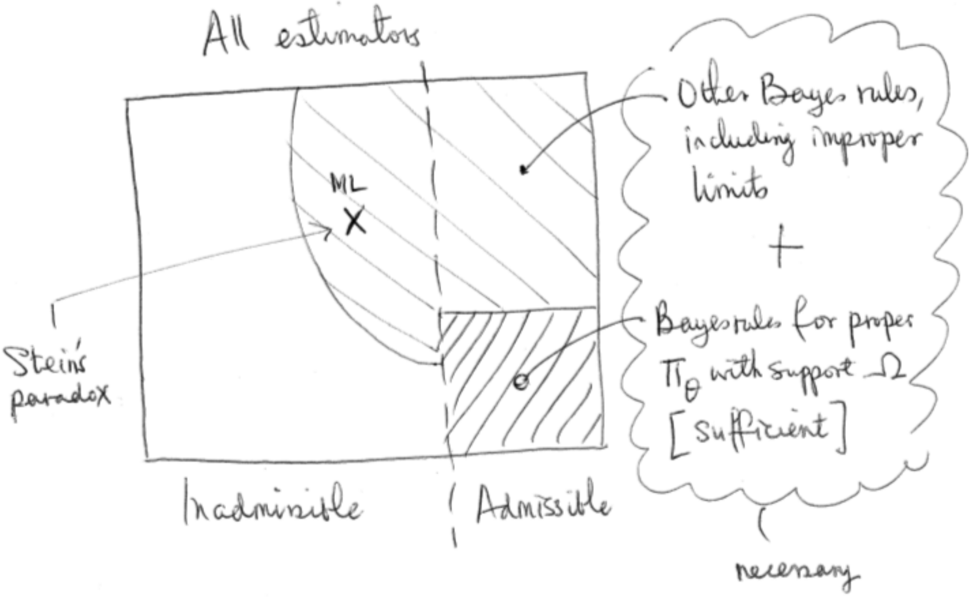
Now consider two different predictors, δ and δ'. We say that δ' dominates δ exactly when

$$R(\delta, \theta) \geq R(\delta', \theta) \text{ for all } \theta \in \Omega$$

with a strict inequality for at least one $\theta \in \Omega$. If δ is not dominated by another rule we say it is admissible. [I did not use finiteness of Ω in this answer.]

(e) State, informally, Wald's theorem. Using a diagram, describe the classification of the space of all estimators, in terms of necessary and sufficient conditions for admissibility. Where does Stein's paradox locate the Maximum Likelihood (ML) estimator in this space?

Answer. It is easy to show that if δ is a Bayes prediction for some π_θ with support Ω then δ is admissible. This is a sufficient condition for admissibility. Wald proved a converse result: all admissible rules are either Bayes rules for a proper π_θ, or, in the case where Ω is non-finite, the improper limit of a sequence of Bayes rules. This is a necessary condition for admissibility. Here is the Venn Diagram of Admissibility, with the ML estimator shown in the necessary set, but not in the admissible set, as per Stein's paradox.


All estimators
ML
X
Stein's paradox
Other Bayes rules, including improper limits
+
Bayes rules for proper Π_θ with support Ω
[sufficient]
necessary
Inadmissible
Admissible

More on Stein's paradox. This comment inserted after looking at the handed-in homeworks. In some cases, including some common ones, the MLE is the improper limit of a sequence of Bayes rules for quadratic loss. So in these situations it satisfies the necessary condition for admissibility. Stein showed that one such MLE, again a common one, was inadmissible.

Please hand in your answers for marking next Tue (17 Mar), at the lecture or by 5pm in the box outside my office door. I will return them in Thu's lecture.

Jonathan Rougier
Mar 2015