Exercises set 6 for the lectures on:
Causality, Boosting and Continuous Latent Variable Models

Due date: 20 July 2019, 17:00

1 Causality (10 points)

Assume that a population of patients contains a fraction r of individuals who suffer from a certain fatal syndrome Z, which simultaneously makes it uncomfortable for them to take a life-prolonging drug X (Figure 1). Let $Z = z_1$ and $Z = z_0$ represent, respectively, the presence and absence of the syndrome, $Y = y_1$ and $Y = y_0$ represent death and survival, respectively, and $X = x_1$ and $X = x_0$ represent taking and not taking the drug. Assume that patients not carrying the syndrome, $Z = z_0$, die with probability p_2 if they take the drug and with probability p_1 if they don't. Patients carrying the syndrome, $Z = z_1$, on the other hand, die with probability p_3 if they do not take the drug and with probability p_4 if they do take the drug. Further, patients having the syndrome are more likely to avoid the drug, with probabilities $q_1 = P(x_1|z_0)$ and $q_2 = P(x_1|z_1)$.

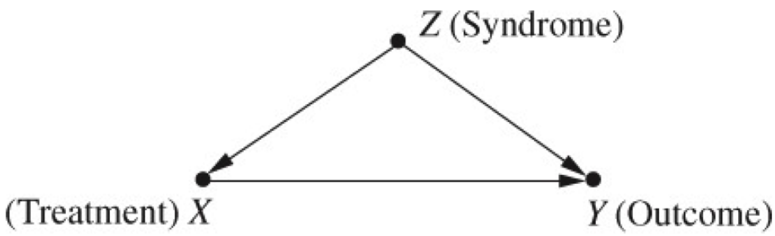

Figure 1: Graphical model for the causality exercises.

E1. (**3p**) Based on this model, compute the joint distributions $P(x, y, z)$, $P(x, y)$, $P(x, z)$ and $P(y, z)$, for all values of x, y and z, in terms of the parameters r, p_1, p_2, p_3, p_4, q_1 and q_2.

E2. (**3p**) Calculate the difference $P(y_1|x_1) - P(y_1|x_0)$ for three populations: (1) those carrying the syndrome, (2) those not carrying the syndrome, and (3) the population as a whole.

E3. (**2p**) Using your results from the previous exercise, find a combination of parameters that exhibits Simpson's reversal.

E4. (**1p**) Compute $p(y|do(x))$ for all possible values of x and y using the adjustment formula.

E5. (**1p**) Compute the Average Causal Effect $ACE = p(y_1|do(x_1)) - p(y_1|do(x_0))$ and compare it to the Risk Difference $RD = p(y_1|x_1) - p(y_1|x_0)$. What is the difference between the two?

2 Boosting (8 points)

The exponential error function of AdaBoost is defined by:

$$E = \sum_{n=1}^{N} \exp(-t_n f_m(x_n))$$

with $f_m(x_n)$ a classifier defined in terms of a linear combination of base classifiers $y_l(x)$ of the form:

$$f_m(x) = \sum_{l=1}^{m} \alpha_l y_l(x)$$

E6. (**3p**) Show that the exponential error function of AdaBoost does not correspond to the log likelihood of any well-behaved probabilistic model. This can be done by showing that the corresponding conditional distribution $p(t|x)$ cannot be correctly normalized.

By a stage wise decomposition the error function E for AdaBoost can be written as:

$$E(\alpha_m) = \left(e^{\alpha_m/2} - e^{-\alpha_m/2}\right) \sum_{n=1}^{N} w_n^{(m)} I(y_m(x_n) \neq t_m) + e^{-\alpha_m/2} \sum_{n=1}^{N} w_n^{(m)}$$

E7. (**2p**) Show by differentiation of the error function wrt α_m that the optimal value of α_m leads to:

$$\alpha_m = \log \frac{1-\epsilon_m}{\epsilon_m}$$

with ϵ_m defined by:

$$\epsilon_m = \frac{\sum_{n=1}^{N} w_n^{(m)} I(y_m(x_n) \neq t_m)}{\sum_{n=1}^{N} w_n^{(m)}}$$

Consider a data set comprising 400 data points from class C_1 and 400 data points from class C_2. Suppose that a tree model A splits these into (300, 100) assigned to the first leaf node (predicting C_1) and (100, 300) assigned to the second leaf node (predicting C_2), where (n, m) denotes that n points come from class C_1 and m points come from class C_2. Similarly, suppose that a second tree model B splits them into (200, 400) and (200, 0), respectively.

E8. (**1p**) Calculate the misclassification rates of the two trees and show that they are equal.

E9. (**2p**) Calculate the cross-entropy and the Gini Index of the two trees and show that tree model B, performs better with both of these scores.

3 Principal Components Analysis and Factor Analysis (4 points)

Write a computer program to generate 200 observations from the following model, where $\mathbf{z} = \mathcal{N}(\mathbf{0}, \mathbf{I}_3)$:

$$\begin{aligned} x_{n1} &= z_{n1} \\ x_{n2} &= z_{n2} + 0.001 z_{n1} \\ x_{n3} &= 10 z_{n3} \end{aligned}$$

with $n \in (1, \ldots, 200)$.

E10. (**4p**) Fit a PCA and an FA model with 1 latent factor. Show that the corresponding weight vector $\mathbf{w}$ aligns with the maximal variance direction (dimension 3) in the PCA case, but with the maximal correlation direction (dimensions 1+2) in the case of FA.