PSM Spring 2019, Exercise sheet 6 - with solutions, for home study

Note. After the quiz 2, and before quiz 3 and the final exam, it appears advisable to train more about core probability concepts in a broad range. The following problems are thus scattered across themes from the entire course up to now, with no chronological order (because I wrote them as they came to my mind). I would consider them all as simple. Solutions are given at the end. For best training effect, try to really fight down each problem before peeking at the solution. The last problem in this set is particularly recommended.

1. Let $(X_n)_{n=1,2,3,4,5}$ be a discrete-time stochastic process whose paths all have length 5. Let the X_n be i.i.d. with values in $\{0,1\}$, where $P_{X_n}(\{0\}) = 0.1$. Let Y_n (where $n = 1, ..., 5$) be defined by $Y_n = X_1 + ... + X_n$. Your tasks: (a) Give the smallest possible sample space S_n for Y_n. (b) Draw a graphics that displays two realizations of the process $(Y_n)_{n=1,2,3,4,5}$.

2. Let $\Omega = \{\omega_1, ..., \omega_5\}$ be a (uncommonly small) universe. Define a probability measure P on Ω and two RVs $X, Y : \Omega \to \{\text{red, blue, green}\}$ which are identically distributed but not identical.

3. Let a three-state homogeneous Markov Chain with states $\{a, b, c\}$ be given by the transition matrix M (whose rows sum to 1). Express the probabilities (a) $P(X_2 = b \mid X_0 = a)$ and (b) $P(X_2 = b \mid X_0 = a, X_1 = c)$ in terms of M.

4. A clinical survey on cancer patients reports that 95% of the patients with blood cancer had antibodies of a certain type A in their blood a year before the cancer became manifest, but from the members of a healthy control that was monitored for 1 year, only 10% carried the antibody A at the beginning of the observation year. Furthermore, it is since long known in the relevant medical literature that the chance for anybody to develop this cancer within 1 year is 0.01%. A doctor who know about all these findings sees a new patient Z of whom tests reveal that he carries antibody A.

   (a) The clinical survey actually describes a statistical model with an underlying probability space $(\Omega, \mathfrak{A}, P)$, a sample space $(E, \mathcal{F})$ and one or several random variables X. Give an informal description of Ω (choose one possibility and detail it out), of E and the random variables X.

   (b) What is the statistical chance of patient Z to develop blood cancer within one year?

5. (*concerning Bayesian model estimation*) Like in the demo example in the LN, we consider a scenario where we are given a sample $X_1(\omega), \ldots, X_N(\omega)$ of a real-valued RV X, and we assume that the true distribution P_X is a normal distribution with unit variance σ^2 but unknown mean μ, that is, we are faced with a 1-parametric family of models where the parameter is $\theta = \mu$. We furthermore assume a prior hyperdistribution for μ that is the normal distribution with zero mean and a standard deviation of 1. Furthermore, we consider a case where the sample size is just $N = 1$ and the (only) sample data point is $X_1(\omega) = -1$. Give a formula for the pdf of the posterior distribution and calculate the posterior mean estimate $\hat{\mu}$ numerical on your computer.

6. At `http://minds.jacobs-university.de/uploads/teaching/PSMSpring2018/FinalMLFall2014_withSolutions.pdf` you find a link to a final exam (with solutions) from a bygone advanced course on ML. This exam consists in a step-by-step discussion of learning a so-called *trigram model* of English texts. Such trigram models are in widespread use in text data mining applications. Much of the themes and methods exhibited in that final exam are accessible to you and may be of interest for you - just "good to know" stuff. I recommend to go through this exam (skipping the tasks nr. 3 and 8 because they relate to methods that we did not cover in class) and try to understand the general set-up, the tasks specifications, and try to find answers yourself before looking at the solutions.

Solution to problem 1. (a) Because the X process can take values 0 or 1 at all times, any natural number between 0 and n is a possible value for Y_n. For Y_n this set is $S_n = \{0, \ldots, n\}$. (b) Any length-5 sequence $s_1, \ldots, s_n$ of natural numbers starting with 0 or 1, which from time n to time $n+1$ either remains at the same level (that is $s_{n+1} = s_n$) or jumps up by 1 (that is $s_{n+1} = s_n + 1$) is a possible realization. I am too lazy to draw two such lines in latex.

Solution to problem 2. Many possibilities. The easiest is to make P the uniform measure on Ω, that is $P(\{\omega_i\}) = 1/5$ for all $i = 1, \ldots, 5$; define X arbitrarily such that X takes at least two values (for example, $X(\omega_1) =$ red and $X(\omega_i) =$ blue for $i > 1$), then permute the i indices by any permutation π which switches two colors (for example, $\pi(1) = 2, \pi(2) = 1, \pi(i) = i$ for $i > 2$), then put $Y(\omega_i) = X(\omega_{\pi(i)})$. In our example this would make $Y(\omega_1) =$ blue, $Y(\omega_2) =$ red, $Y(\omega_i) =$ blue for $i > 2$.

Solution to problem 3. Let e_i (where $i = 1, 2, 3$) be the ith unit vector, that is $e_1 = (100)', e_2 = (010)', e_3 = (001)'$. Denote, for a vector w, the ith vector component by $w[i]$. Then (a) $P(X_2 = b \mid X_0 = a) = ((M')^2\, e_1)[2]$ and (b) $P(X_2 = b \mid X_0 = a, X_1 = c) = (M'\, e_3)[2] = M_{3,2}$.

Solution to problem 4. (a) Ω: For instance, take for Ω the set of all people in the country where the survey was carried out. However, this is likely not fully appropriate. Clinical studies typically don't pick people fully at random. The probability of developing cancer given antibody A reported in the study does not really reflect a population probability, but likely (that's how such studies go...) a probability describing a subpopulation of people that happened to go to hospital for one or the other reason... (and there are recruited for the survey). If the study would be transparent and honest, it should be stated that Ω is the subset of citizens that become hospitalized... or something to that effect. There are two random variables involved, both binary, that is, $E = \{0, 1\}$: X indicates whether a patient develops cancer within one year, Y indicates whether a patient carries antibody A.

(b) The doctor wants to know the probability $P(X = 1 \mid Y = 1)$ that Z develops cancer given that he carries antibody A. The clinical survey provides $P(Y = 1 \mid X = 1)$ instead. Bayes' rule can transform this into the desired probability by $P((X = 1 \mid Y = 1) = P(Y = 1 \mid X = 1)P(X = 1)/P(Y = 1)$. $P(X = 1)$ is known from the medical literature. $P(Y = 1)$ can be computed by marginalization: $P(Y = 1) = P(Y = 1, X = 1) + P(Y = 1, X = 0) = P(Y = 1 \mid X = 1)\, P(X = 1) + P(Y = 1 \mid X = 0)\, P(X = 0) = 0.95 * 0.0001 + 0.1 * (10.0001) = 0.1001$. From this one gets $P(X = 1 \mid Y = 1) = 0.000949 = 0.0949\%$.

Solution to problem 5. We have $D = \{-1\}$ and

$$p_X(D \mid \mu) = p_{X_1}(D \mid \mu) = 1/\sqrt{2\pi}\ \exp(-(-1-\mu)^2/2)$$

and

$$h(\mu) = 1/\sqrt{2\pi}\ \exp(-(\mu)^2/2),$$

which gives a posterior distribution

$$\begin{aligned} h(\mu \mid D) &= \frac{1/2\pi\ \exp(-(-1-\mu)^2/2)\ \exp(-(\mu)^2/2)}{\int_{\mathbb{R}} 1/2\pi\ \exp(-(-1-\mu)^2/2)\ \exp(-(\mu)^2/2)\ d\mu} \\ &= \frac{\exp(-\frac{1+\mu+\mu^2}{2})}{\int_{\mathbb{R}} \exp(-\frac{1+\mu+\mu^2}{2})\ d\mu}. \end{aligned}$$

To get the posterior mean estimate, we numerically calculate

$$\hat{\mu} = \int_{\mathbb{R}} \mu\ h(\mu \mid D)\ d\mu \approx -0.5.$$