Machine Learning, Spring 2017: Exercise Sheet 1

Please send your type-set solutions by email to our two TA's Xu He ("Owen") x.he@jacobs-university.de and Felix Schmoll f.schmoll@jacobs-university.de. Join into groups of two or three and submit a single solution sheet per group, indicating the group members' names on the sheet.

Deadline for submission is Friday Feb 17, 23:59 hrs (email sending timestamp). Submissions arriving later (even a second after midnight) will be corrected but not counted for the course grade.

Note: you'll need to know the material from the Appendix A in the lecture notes to work out these exercises.

Problem 1 (easy warm-up) In the lecture notes I (Section 2) I was describing the TICS system with the use of two random variables which I called Image and Caption. Possible values of the Image RV were all vectors in $[0, 1]^{144000}$, and of Caption, any word sequence of length at most 20 made from a given vocabulary. Generally speaking, a RV always comes with a set of technically possible values, called the *measure space* of that RV. In order to get more familiar with this concept of a RV and its associated measure space, consider a scientific study carried out by an experimental psychologist where subjects first listened to a recording of a series of 10 randomly picked numbers (between 0 and 9) spoken by a synthetic voice, then had to wait 5 seconds, then had to repeat as much as they could remember from that sequence, starting from the sequence's beginning in the right order. Per subject, this trial was repeated 20 times with freshly randomly picked number sequences. In some of the trials, the voice was speaking in a neutral tone, in other trials the voice was speaking in an excited tone. The purpose of the experiment was to find out to what extent (if any) information given by an excited speaker is better remembered than information from a neutral speaker. Your task: declare a choice of random variables, each one describing some aspect of a trial, such that together they convey all the information that is relevant for this study. For each RV that you think of, (i) give a brief description in plain English of what it measures, (ii) give it a telling name, (iii) describe its measure space, (iv) state whether it is discrete or continuous. – The relevant information in these trials can be captured by observations (that is, RVs) in various ways, there is not a uniquely correct choice of them. Give at least five RVs.

Solution (sketch). Two examples

1. Description: number of correctly recalled initial sequence items. Name: Ncorrect. Measure space: {0, 1, ..., 10}. Type: discrete.

2. Description: time after go signal to uttering first answer. Name: Tstart. Measure space: $\mathbb{R}^{\geq 0}$.

Problem 2 (some work and some thinking). In a clinical study a population of male 1540 patients was tested for the number of erythrocytes (red blood cells) per cubic millimeter of blood. In healthy adult male humans one finds about 5 – 6 million erythrocytes / ml. Since this counting is not very precise, the data were binned in bins

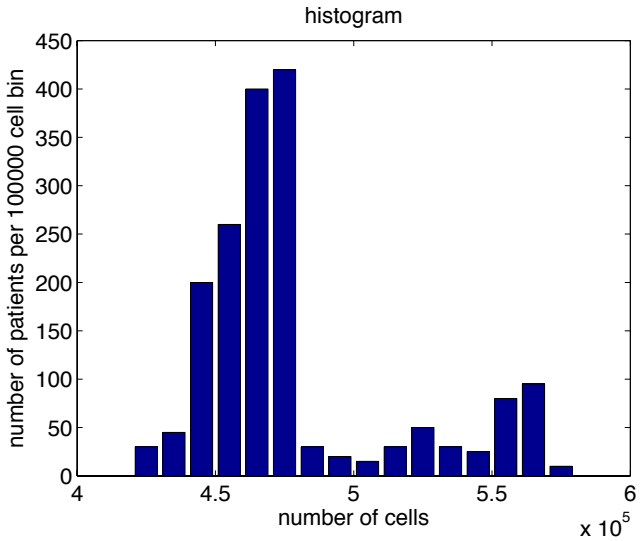
of width 100,000. The obtained counts are given by the count vector (a *histogram*) shown in the following table (first row: Millions of cells, second row: nr of patients

4.0-4.1	4.1-4.2	4.2-4.3	4.3-4.4	4.4-4.5	4.5-4.6	4.6-4.7	4.7-4.8	4.8-4.9	4.9-5.0
0	0	30	45	200	260	400	420	30	20

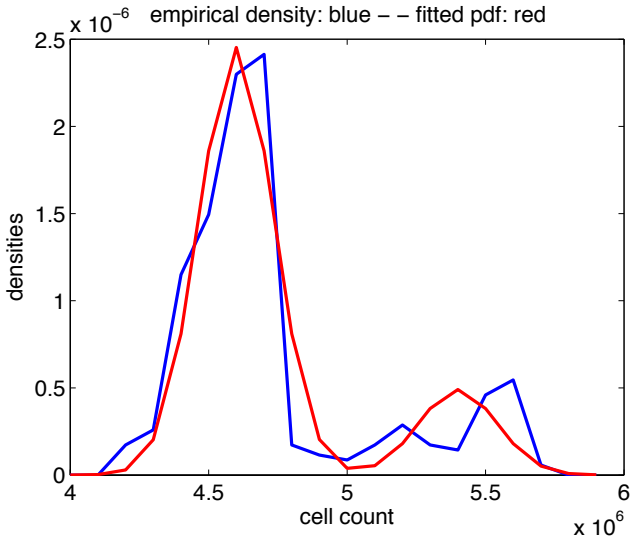
5.0-5.1	5.1-5.2	5.2-5.3	5.3-5.4	5.4-5.5	5.5-5.6	5.6-5.7	5.7-5.8	5.8-5.9	5.9-6.0
15	30	50	30	25	80	95	10	0	0

Your task: use Matlab or Python to calculate and plot a smooth pfd over the interval [4*10E6, 6*10E6], which captures the essential shape of this histogram. For this task, you have to familiarize yourself with interpolation methods offered by Matlab or Python (in the scipy or numpy packages), and with plotting routines. You'll need these things in other exercises forthcoming in this course. Besides the plot, provide a brief documentation of how you derived the pdf from the histogram.

Solution. Here is how your solutions should look like:



This first figure depicts the histogram of the given data. Not required in the task description but instructive to see. Notice that by definition a *histogram* shows the raw counts, without normalization. The next picture shows what your solution should look like:



The second picture shows two lines. The blue one plots the original cell counts in a density format. Notice the small values on the y axis: the max value shown is only about 2.5 * 10e-6. This makes the integral under the curve to take a value of 1! The red line shows a hand-made smoother fit to the blue line obtained from a weighted sum of two Gaussians. It also is normalized to integrate to 1. If one wishes an even smoother looking curve, one would have to use a finer plotting resolution and interpolate the red (or blue) curve with some smooting filter (Matlab offers many, e.g. spline interpolation). At any rate, your solution curves should not have negative values, integrate to 1, and be essentially zero at the left and right ends of the picture.

Problem 3 (very easy). Here we use the standard shorthand notation $P(x, y)$ for $P(X = x, Y = y)$, $P(x \mid y)$ for $P(X = x \mid Y = y)$, etc. Prove the following *factorization formula*

(1) $\quad P(x, y, z) = P(x)\, P(y \mid x)\, P(z \mid x, y),$

starting from the definition of conditional (discrete) probabilities

(2) $\quad P(u \mid v_1, \ldots, v_n) = P(u, v_1, \ldots, v_n) \,/\, P(v_1, \ldots, v_n).$

Note: in statistics and machine learning in general, factorizing joint distributions into products of simpler distributions is a super common strategy.

Solution. Using (2), rewrite the rhs of (1) as

$$P(x)\, P(y \mid x)\, P(z \mid x, y) = P(x) \frac{P(x,y)}{P(x)} \frac{P(x,y,z)}{P(x,y)} = P(x, y, z).$$