TMS150 / MSG400
Stochastic data processing and simulation
Autumn 2020
Lab 4: Bayesian decision theory

Petter Mostad

October 1, 2020

These are lecture notes for Lab 4 of the 2020 version of TMS150/MSG400. Some parts are built on an eariler version written by Patrik Albin.

1 Bayesian inference

Assume X and Y are random events. Then according to Bayes formula,

$$\Pr(X \mid Y) = \frac{\Pr(Y \mid X)\Pr(X)}{\Pr(Y)}.$$

Bayes formula also holds for densities and probability mass functions:

$$\pi(x \mid y) = \frac{\pi(y \mid x)\pi(x)}{\pi(y)} \tag{1}$$

where we have a joint density on variables x and y, and a generic π notation is used, so that for example $\pi(x)$ is the density value at x for the marginal density of x, and $\pi(y \mid x)$ is the density value at y of the conditional density of y given x.

In computational statistics, one may often approximate a continuous density with a discrete density: If for example a real variable x with $0 < x < 1$ has a continuous density $f(x)$, we may approximate it by choosing equally spaced values $x_1, \ldots, x_n$ in the interval $(0, 1)$ and setting the probability for each such value, denoted $\Pr(x_i)$ or $\pi(x_i)$, to

$$\pi(x_i) = \frac{f(x_i)}{\sum_{j=1}^{n} f(x_j)}.$$

More generally, if $f(x, y)$ is a bivariate density on the variables $L_1 < x < H_1$ and $L_2 < y < H_2$ we may approximate it by choosing a grid of points (x_i, y_j),

with $i = 1, \ldots, n$ and $j = 1, \ldots, m$ and computing probabilities

$$\pi(x_i, y_j) = \frac{f(x_i, y_j)}{\sum_{k=1}^{n} \sum_{s=1}^{m} f(x_k, y_s)}.$$

In this text, we will approximate all continuous densities with discrete probability mass functions as above.

For discrete probaility mass functions, marginal mass functions are computed as sums. So Bayes formula can be written

$$\pi(x_i \mid y_j) = \frac{\pi(y_j \mid x_i)\pi(x_i)}{\pi(y_j)} = \frac{\pi(y_j \mid x_i)\pi(x_i)}{\sum_{k=1}^{m} \pi(x_k, y_j)} = \frac{\pi(y_j \mid x_i)\pi(x_i)}{\sum_{k=1}^{m} \pi(y_j \mid x_k)\pi(x_k)}.$$

As an example, assume y has a Binomial distribution with parameters n and x (with $0 < x < 1$) written $y \sim \text{Binomial}(n, x)$, so that y is the number of "successes" among n independent trials when the probability of "success" in each trial is x. Then we have

$$\pi(y \mid x) = \binom{n}{y} x^y (1 - x)^{n-y}. \tag{2}$$

Assume y is observed and we want to use that observation to learn about x. In Bayesian staitistics, one always assumes there is some *prior knowledge* about x.

In our case such prior knowledge can be formulated as a probability density on the interval $(0, 1)$, or a probability mass function for values inside this interval. Assume, for example, we have the discrete prior illustrated in Figure 1. The *posterior distribution* for x is the conditional distribution $x \mid y$. Using Equation 1, we can compute $\pi(x \mid y)$ by computing $\pi(x)$ for each of the x for which it is non-zero (see Figure 1), and multiply with $\pi(y \mid x)$ computed from Equation 2 at the same values. We do not need to explicitly compute $\pi(y)$, as we can instead normalize the vector of products so that it sums to 1. Assuming $y = 16$ and $n = 31$, Figure 2 shows the prior $\pi(x)$, the likelihood $\pi(y \mid x)$ and the posterior $\pi(x \mid y)$.

Having put what we have learned about x from y in the posterior $\pi(x \mid y)$, the most important use for this posterior is to make predictions for new observations, taking into account the uncertainty in the parameter x. In general, in classical statistics one uses y to find an *estimate* $\hat{x}$, and then predict new observations y_{new} with the distribution $\pi(y_{new} \mid \hat{x})$. In Bayesian statistics one *integrates out* the uncertainty in the parameter x, using for prediction

$$\pi(y_{new} \mid y) = \int \pi(y_{new} \mid x)\pi(x \mid y)\, dx.$$

When x is a discrete variable, the integral becomes a sum:

$$\pi(y_{new} \mid y) = \sum_{\text{all } x} \pi(y_{new} \mid x)\pi(x \mid y) \tag{3}$$

In the previous example, assume we would like to predict the number of successes in 10 new trials. In a classical analysis one would first use maximum

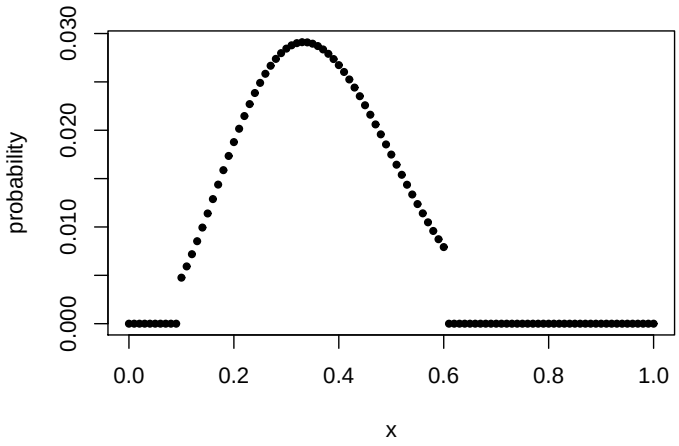

Figure 1: Prior for x for Example 1: The plot shows the probabilities $\Pr(x = 0)$, $\Pr(x = 0.01)$, $\Pr(x = 0.02)$, ..., $\Pr(x = 0.99)$, and $\Pr(x = 1)$.

likelihood to estimate $\hat{x} = 16/31 = 0.5161$ from the data. Using the Binomial distribution with parameters 10 and 0.5161 one would compute the predictive probabilities shown as black dots in Figure 3. In a Bayesian anlaysis, one would instead use the posterior obtained above, and Equation 3, replacing $\pi(y_{new} \mid x)$ with the Binomial distribution with parameters 10 and x, to obtain the predictive probabilities shown in red triangles.

We also illuatrate a 2D example of Bayesian inference. Assume that $x = (x_1, x_2)$, with $0 \leq x_1 \leq 1$ and $0 \leq x_2 \leq 1$, and that we have a prior on x as illustrated in Figure 4. Note how the plots in this figure are generated with the useful `image` R command. The prior is actually a discrete distribution on the nodes of a grid of size 101×101 covering the rectangle above. Figure 4 also shows a likelihood density $\pi(y \mid x)$: It is actually a bivariate normal distribution with expectation equal to x. Note that if you have programmed in R a function that computes the likelihood based on arguments `x1` and `x2`, the R function `outer` may be very useful to obtain a matrix of likelihood values like the one shown Figure 4. Finally, the figure shows the resulting posterior distribution, which has been obtained by pointwise multiplication of the prior and the likelihood, followed by normalization so that the 10201 probabilities of the 101×101 grid sum to 1.

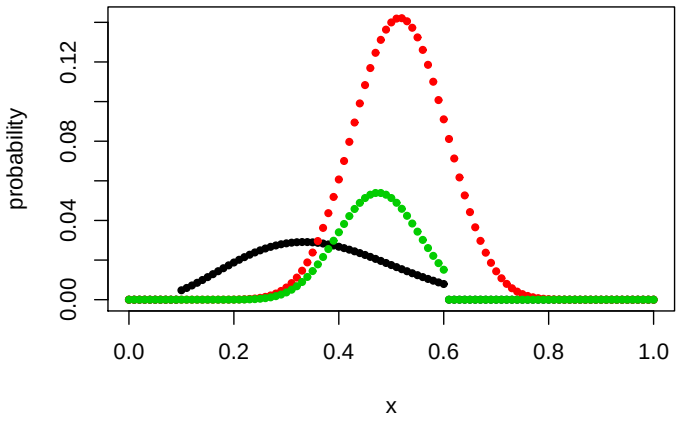

Figure 2: The prior $\pi(x)$ for x from Figure 1 is shown in black; the likelihood $\pi(y \mid x)$ from Equation2 with $y = 16$ and $n = 31$ is shown in red, and the posterior $\pi(x \mid y)$ is shown in green.

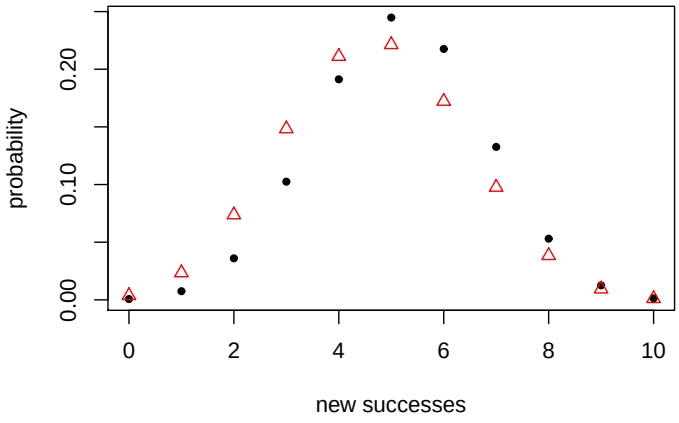

Figure 3: In black dots: The predictions for the number of "successes" in 10 new trials in a classical analysis. In red triangles are the predictions using a Bayesian analysis.

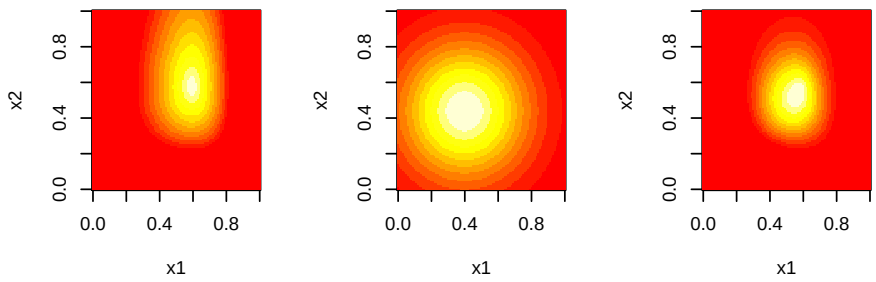

Figure 4: The prior (left), likelihood (center), and posterior (right) densities of a 2D example.

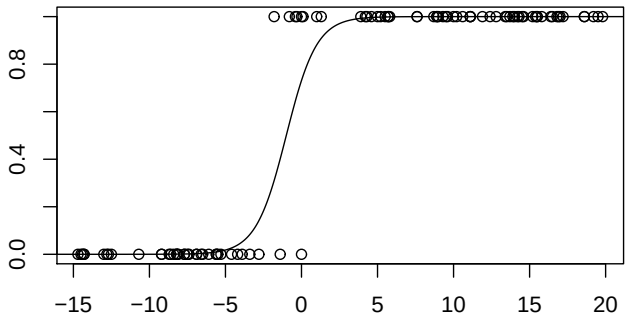

Figure 5: An example using logistic regression. The data is shown together with a possible logistic regression curve.

2 Logistic regression

Figure 5 shows some (synthetic) data from a drug trial: The x-axis is a measure of the drug concentration that each test animal has received, and the y-axis indicates a response 1 for each animal that has shown an adverse reaction, and a response 0 otherwise. Clearly, the probability for an adverse reaction depends on the drug concentration x. For an animal with drug concentration x and

response y, a natural model may be

$$y \sim \text{Bernoulli}(p(x)) \quad \text{and} \quad p(x) = \frac{\exp(a+bx)}{1+\exp(a+bx)},$$

where a and b are constants. Note that $y \sim \text{Bernoulli}(p(x))$ means that y is 1 with probability $p(x)$, otherwise y is 0. As an example, Figure 5 illustrates $p(x)$ when $a = 1$ and $b = 1$.

Assume now that we have observed the data illustrated in Figure 5 and that we would like to make Bayesian inference for the model parametres a and b. Experience with earlier drug testing might provide the information that $0 < a < 5$ and $0.1 < b < 3$. For simplicity, we use a prior that is uniform on the 101×101 grid on the rectangle $[0, 5] \times [0.1, 3]$. If we denote the data with $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the probability of this data given a and b, i.e., the likelihood function, can be computed as

$$L(a,b) = \prod_{i=1}^{n} \left(\frac{\exp(a+bx_i)}{1+\exp(a+bx_i)} \right)^{y_i} \left(1 - \frac{\exp(a+bx_i)}{1+\exp(a+bx_i)} \right)^{1-y_i}$$

The resulting posterior density is shown in Figure 6.

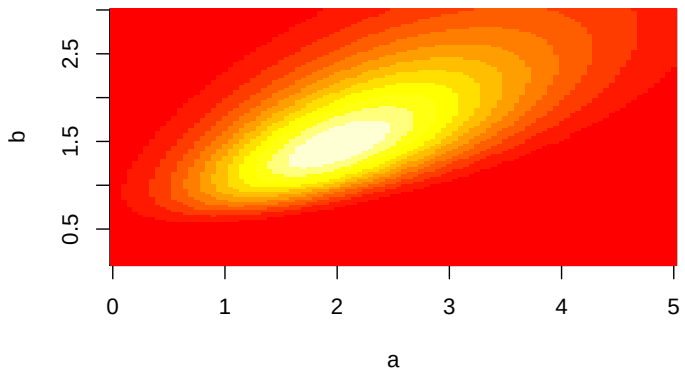

Figure 6: The posterior over the parameters a and b in the logistic regression.

3 Decision theory

Decision theory is a structured way to obtain optimal decisions when decisions have to be taken under uncertainty. More precisely, we assume that we are to select from set of *actions* $\mathcal{A}$; this set may be finite or infinite. The outcome of

an action $a \in \mathcal{A}$ will depend on the *state of nature*; let us use Θ to denote all the possible relevant states of nature. This is where the uncertainty comes in; we assume we have a probability distribution on Θ describing our knowledge about it. We also assume that we can describe a function f which to any action $a \in \mathcal{A}$ and any state of nature $\theta \in \Theta$ describes the outcome $f(a, \theta)$ within some relevant set of outcomes.

In order to compare actions, we need to compare the possible outcomes, so we assume we have a *utility function* $u(x)$ which to any outcome x assigns a real value, the *utility* of the outcome. The utility is meant to measure the usefulness or desireability of the outcome. Sometimes it is natural to talk about the negative utility, which is then termed *cost*. The main principle of decision theory says that we should choose the action $a \in \mathcal{A}$ which maximizes the expected utility, i.e., $\mathrm{E}[u(f(a, \theta))]$, where the expectation is taken over over probability distribution for θ.

In the stock optimization example below, the probability distribution for θ is derived from data using some classical statistics. In the following medical age assessment example, the probability distribution for θ is derived as a posterior distribution using Bayesian statistics. Thus in this second example, one would talk about Bayesian decision theory.

Clearly, in order to make good decisions, we always need to compare how useful or desireable various outcomes are to us. Nevertheless, specifying a precise utility function can be a challenge. The easiest context to work in may be when the outcome of our action is a monetary profit or loss, because we immediately have at least a ranking of such outcomes. However, it is not necessarily true that the utility should be a linear function of the amount of money we win or loose.

Consider the following utility function for positive x:

$$u(x) = \frac{1 - (x/K)^{-k}}{k}$$

where $k \neq 0$ is some parameter and K is some amount of money that we invest. The function shows the (subjective) utility of being left with the monetary amount x. Figure 7 shows $u(x)$ for a range of positive k values. As can be easily shown, we always have $u(K) = 0$ and $u'(K) = 1$. In fact, if we set $u(x) = \log(x/K)$ when $k = 0$ we obtain a family of utility functions valid for any real k. When $k = -1$ the utility of the gained or lost money is proportional to the amount gained or lost, but as k increases, risk aversion becomes more and more apparent: Money gained is of some but not much utility, while money lost decreases the utility a lot. This may be a good model for people for whom the most important thing is to avoid loosing all or most of K. However, there are also *risk seeking* people who have the opposite attitude: The most important thing is to have a shot at the big prize, how much is lost is less important. The utility of such people might also be modelled with the function above, but now with $k < -1$.

If we assume that we know the function f relating each action and each state of nature to an outcome, and that we can describe our knowledge about θ in a

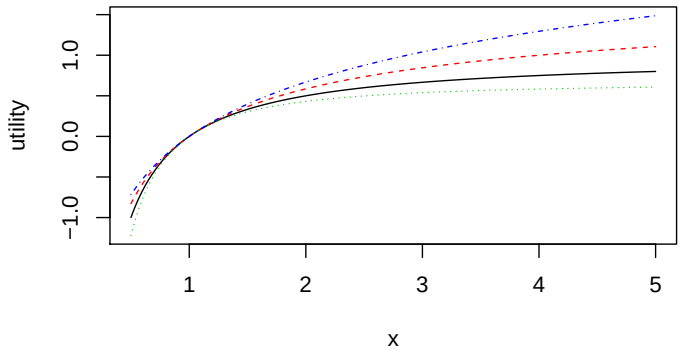

Figure 7: Utility functions of the form presented in Example 3, with different values for k. The values $k = 1.5, 1, 0.5, 0.1$ are represented by the green dotted line, the full line, the red dotted line, and the blue dotted line, respectively. The x-axis is in terms of multiples of K.

probability distribution, the challenge of decision theory is to find the action a mazimizing $\mathrm{E}[u(f(a, \theta))]$. Exactly how this is done depends on the model. We will look at two quite different examples.

4 Example: Stock optimization

Predicting how the stock market moves is not easy. Here, we will consider a simple but useful way of modelling how the daily closing prices of a set of stocks develop over time. Let X_{ij} be the closing price of stock i ($i = 1, \ldots, m$) after day j ($j = 0, \ldots, N$). First of all, the relevant way to measure changes in stock prices is to look at relative changes, i.e., percentage changes. Thus we define, for $i = 1, \ldots, m$ and $j = 1, \ldots, N$,

$$Z_{ij} = \log\left(\frac{X_{ij}}{X_{i,j-1}}\right) = \log(X_{ij}) - \log(X_{i,j-1}).$$

In our simple model, we will assume that there are underlying trends in the prices of the stocks, and that some stocks move together, but apart from this, the day-to-day changes are random and independent. A reasonable model then turns out to be a multivariate normal model where the vectors $Z_j = (Z_{1j}, Z_{2j}, \ldots, Z_{mj})$ are independent for different j's and we have

$$Z_j \sim \text{Normal}\,(\gamma, \Sigma)$$

for some vector γ and some covariance matrix Σ. In other words, the Z_j are a sample from a multivariate normal distribution with expectation γ and covariance matrix Σ. Note how γ and Σ may be estimated from data using the mean and sample covariance matrix for the differences in logtransformed data.

Assume we want to invest a total amount K in these stocks, with each stock i receiving a proportion w_i of the investment. Then the logged value of stock i is initially $\log(Kw_i)$, and each of n days of investment this logged amount is increased with a random value according to the model above. The total amount after n days becomes

$$T = \sum_{i=1}^{m} \exp(\log(Kw_i) + Y_i) = K \sum_{i=1}^{m} w_i \exp(Y_i)$$

where

$$Y = (Y_1, Y_2, \ldots, Y_m) \sim \text{Normal}\left(n\gamma, n\Sigma\right). \tag{4}$$

Assume we will use decision theory to obtain a set of weights $w = (w_1, \ldots, w_k)$ that optimizes the utility after n days. Assume that we use the utility function introduced in Example 3 with a starting value of K. The utility after n days becomes

$$u(T) \quad = \quad u\left(K \sum_{i=1}^{m} w_i \exp(Y_i)\right) = \frac{1 - (\sum_{i=1}^{m} w_i \exp(Y_i))^{-k}}{k}$$

THIS FORMULA WAS WRONG IN THE PREVIOUS VERSION OF THESE NOTES.

We would like to compute the expecation of this utility under the distribution of Equation 4. Writing this down as an iterated integral will create something that is difficult to compute. We will here consider an approximate solution where we compute an expectation by averaging over a sample from the relevant distribution.

Assume we generate a sample $V_1, \ldots, V_S$ of S vectors each of length m, where for each $V_q = (V_{q1}, V_{q2}, \ldots, V_{qm})$ we have

$$V_q \sim \text{Normal}\left(n\gamma, n\Sigma\right). \tag{5}$$

Then we may approximate

$$\text{E}(u(T)) \approx \frac{1}{S} \sum_{q=1}^{S} \frac{1}{k} \left(1 - \left(\sum_{i=1}^{m} w_i \exp(V_{qi})\right)^{-k}\right). \tag{6}$$

THIS FORMULA WAS WRONG IN THE PREVIOUS VERSION OF THESE NOTES.

A sample like this may be generated using for example the R function `rmnorm` of the R package `LearnBayes`.

5 Example: Medical age assessment

To illustrate a quite different appliation of decision theory, we now consider *medical age assessment*, i.e., using observed biological features, such as the maturity of teeth, bones, etc., to assess a person's age. This may be important for example in legal contexts, where laws regarding migration or sentencing for crimes are different depending on whether the subject is above or below 18 years. The biological features that are observed can include the root systems of wisdom teeth, the growth zones of various bones in the wrist, knee, or shoulder, or even molecular data from cells. Each feature may be classified according to a graded scale, and information from various features may then be combined. For simplicity, we will assume that the final medical report can have one of K different states. For each of these states, there is then a probability that this state will be observed at a given age x. We will, initially, assume that functions $f_k(x)$ are known ($k = 1, \ldots, K$), where $f_k(x)$ specifies the probability that medical observations will result in report state k when applied to a person of age x. Note that we have $\sum_{k=1}^{K} f_k(x) = 1$ for all ages x. From now on we will assume that there are only two possible reports, a "mature" report with probability $f(x)$ and an "immature" report with probability $1 - f(x)$, when the true age is x.

One biological feature which is sometimes used in age assessment is the maturity of teeth (more precisely the developmental stage of the roots of third molars). If we make the simplification that we only look at whether the teeth have reached the final stage or not, we may model the probability for such maturity using logistic regression. A possibility is then two use the model

$$y \sim \text{Bernoulli}(f(x)) \quad \text{and} \quad f(x) = \frac{\exp(a + b(x - 18))}{1 + \exp(a + b(x - 18)}.$$

Here y is 1 if the teeth are mature and 0 if not, while x is the age of the person. Note that we have adjusted the formula for $f(x)$ by subtracting 18 from x: Such an adjustment decreases the dependency between a and b and improves the numerical properties of inference procedures.

Normally, a legal determination of age should be based on many sources of information; in our example, we will look at the option of basing a decision solely on the medical report. The legal authority will then have to decide, for each report state, whether persons with this report should be classified as adults (i.e., over 18) or children. It is the optimal selection of this decision rule that we will analyze using decision theory.

An important ingredient will then be the distribution of the true ages of the persons to which the medical age assessment procedure is applied. It is of course impossible to determine this distribution with high certainty, as the ages of these persons are indeed in dispute, at least at the time when the age assessment report is produced. However, it can be argued that, even if the uncertainty is wide, some knowledge about the age distribution exists. In our example, we will assume that the true age is at least 14, and that the number of years above 14 is modelled with a Gamma density. Specifically, we will assume

that the density given parameters $\alpha > 0$ and $\mu > 14$ is

$$\begin{aligned}\pi(x;\mu,\alpha) &= \text{Gamma}(x-14;\alpha,\alpha/(\mu-14)) \\ &= \begin{cases} \frac{\left(\frac{\alpha}{\mu-14}\right)^{\alpha}}{\Gamma(\alpha)}(x-14)^{\alpha-1}\exp\left(-\frac{\alpha}{\mu-14}(x-14)\right) & \text{if } x \geq 14 \\ 0 & \text{if } x < 14 \end{cases}\end{aligned}$$

The density has two parameters: μ which gives the average age, and α which is related to the spread around this age. For various choices of μ and α we get age distributions which may seem reasonable. Figure 8 illustrates these densities for some choices of μ and α.

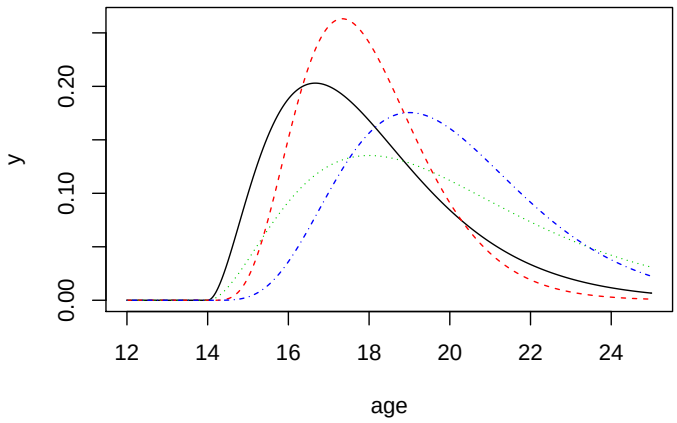

Figure 8: The black and red curves show population distributions with expectation $\mu = 18$ years. The black curve uses $\alpha = 3$ and the red curve $\alpha = 6$. The green and blue curves show population distributions with expectation $\mu = 20$ years. The blue curve uses $\alpha = 3$ and the green curve uses $\alpha = 6$.

In order to do decision theory, we will need some cost or utility functions. It seems most natural to use cost functions here, as the main costs are the "costs" when a child is classified as an adult, and the "costs" when an adult is classified as a child. We will assign zero cost or utility to the cases where a person is classified into the correct age group. Many of the "costs" in this connection are not monetary costs, but may instead be emotional costs, ethical costs, etc. Nontheless, when making a decision, at some point the desireability of all outcomes need to be weighed against each other. In decision theory, this is facilitated by placing the outcomes along a single line of desireability, describing this as measuring their costs (or utilities). Note however that the costs do not need to be denominated in kronor or euro, and that only the relative values of the costs will influence the results.

We will in our example try out two different cost functions: One where the cost of classifying a child as an adult is B times the cost of classifying an adult as a child, so that the cost of misclassification is

$$c_1(x) = \begin{cases} B & \text{if } x \leq 18 \\ 1 & \text{if } x > 18 \end{cases} \tag{7}$$

and one where the misclassification cost is based on the difference between the actual age and 18:

$$c_2(x) = \begin{cases} B(18-x) & \text{if } x \leq 18 \\ x - 18 & \text{if } x > 18 \end{cases} \tag{8}$$

Ethical discussions can be made about the appropriate value for B. For simplicity, from now on we fix $B = 10$.

We now have the ingredients to formulate decision theory for this application: For each of the "mature" report and the "immature" report, one will classify those who receive this report either as children or as adults. If one classifies as children, the expected cost is the cost of misclassification, i.e.,

$$C_c = \int_{18}^{\infty} \pi(x; \mu, \alpha) f_k(x) c(x)\, dx \tag{9}$$

where $c(x)$ is either of the cost functions defined above and $f_k(x)$ is either $f(x)$ or $1 - f(x)$ depending on whether the report is "mature" or "immature" . If one classifes as adults, the expected cost is

$$C_a = \int_{0}^{18} \pi(x; \mu, \alpha) f_k(x) c(x)\, dx. \tag{10}$$

Thus one should classify as children if $C_c < C_a$ and as adults if $C_c > C_a$. The integrals may be computed for example using numerical integration.

6 Questions

The questions should be answered in a full report, of similar type as for Lab 2. The questions below give a total of 11 points. You need at least 6 points to pass this lab.

1. We first consider the stock investment example.

   (a) (1 point) Consider the data `stockvalues.txt` accessible from a link on the syllabus page of Canvas. (Hint: Read the data into R with for example `as.matrix(read.csv("stockvalues.txt"))`). The data lists the stock prices for 7 different stocks each day from 2002-06-03 to 2006-06-01. Assume you want to invest in a combination of these stocks over a period of $n = 100$ days. For this purpose, generate vectors as those of Equation 5, using a simulation size of $S = 1000$.

(b) (1 point) Using weights $w = (w_1, \ldots, w_7)$ (summing to 1) for the different stocks, implement an R function that, given the weights w and the utility function parameter k, computes the approximate expected utility. Compute the value of the function for each of the cases $k = -0.5, k = 0.5$, and $k = 1.5$, and for the two possibilities where each stock receives equal investment, and where the stock with the best expected performance receives all the investment.

(c) (1 point) Assume you limit yourself to investing in the stocks S3 and S4. Use optimization in R to find the optimal weights when $k = -0.5$ and when $k = 1.5$.

(d) (1 point) There are a number of problems and weaknesses with using the approach above to reach a decision about how to invest. These problems can be subdivided into problems regarding the model we use (i.e., the assumptions we make) and problems with how we do the computations. Discuss a number of these problems.

2. Our second task concerns medical age assessment. We consider a single medical feature, maturity of knees (or more precisely, the maturity of the growth zone of the distal femur). Simplifying, a knee can be reported as either "mature" or "immature", so there are only two possible reports. On the syllabus page of Canvas you can find a link to the file matureKnee.txt which lists the ages of 50 people with mature knees and immatureKnee.txt, which lists the ages of 50 people with immature knees.

   (a) (2 points) Fit a logistic regression to the data in the data files by numerically maximizing the likelihood for the data given parameters a and b. Plot the data and the resulting logistic regression curve, and check that your result is reasonable. (Hint: There are pre-programmed maximization algorithms in R. Second hint: If you have numerical problems working with the likelihood function, it may help to work with the logarithm of the likelihood function instead).

   (b) (2 points) Make a grid of size 21×21, where a has evenly spaced values from -0.5 to 2 and b has evenly spaced values from 0.5 to 3. Use a prior that is uniform on this grid. Compute and plot the posterior, visualizing with the *image* plotting function.

   (c) (1.5 points) We now assume that the forensic report is "mature knee". Make a function which, for fixed a and b specifying the logistic report probability function $f(x)$, and for either of the cost functions $c_1(x)$ and $c_2(x)$ defined in Section 5, computes the difference between the cost of classifying as a child and the cost of classifying as an adult. In this question, use the estimates for a and b found in question (a). For the age distribution, you should try out three possibilities:

      i. Agedistribution 1: $\mu = 18.5$, $\alpha = 3$.
      ii. Agedistribution 2: $\mu = 19.5$, $\alpha = 6$.
      iii. Agedistribution 3: $\mu = 20.58$, $\alpha = 3$.

Conclude and list the situations you have found in which a "mature knee" should result in a classification as an adult.

(d) (1.5 points) Finally, we should take into consideration the uncertainty in a and b: Use the discrete distribution on a and b found in (b) and average the results of the function found in (c) over this distribution. These averaged results should be computed for the same cases as in (c). Compare the results, and comment on any differences.