SDS3386 Data Science Lab

Assignment 3. Due Thursday October 13^{th} at 6 pm

2022 Fall, Tanya Schmah

Instructions: This assignment may be done as a group of 2 or 3, or individually – it's your choice. For Questions 1 and 2, you may write your answers by hand, or in LaTeX or Word, or however you like, but somehow you should submit a PDF. For Question 3, you should submit a Jupyter Notebook with your answers, and also export it to HTML and submit that, making 3 files in total. Both the PDF and the Notebook should contain your name and (if the case) the names of your other group members. All group members should submit the identical 3 files to Brightspace.

1. Consider a random screening program for a disease that affects 1 in 500 people in the general population. Suppose the screening test has a *sensitivity* of 99%, meaning that the test is positive 99% of the time in people who actually have the disease; and a *specificity* of 98%, meaning that when people who don't have the disease are tested, the test gives a false positive result 2% of the time. If someone tests positive, what is the probability that they actually have the disease? [2pts]

   Hint: Use Bayes' Rule.

2. A coin, with fixed but unknown probability $\theta \in (0, 1)$ of coming up heads, is flipped N times. Let $(X_1, X_2, \ldots, X_n)$ be the Bernoulli random variables representing the N outcomes, assumed all independent. We observe the results, giving us a dataset $D = (x_1, x_2, \ldots, x_n)$, where $x_i \in \{0, 1\}$ for all i, where $x_i = 1$ means heads, and $x_i = 0$ means tails. The probability of observing the data, given a certain value of θ, is: [3pts]

$$P(D \mid \theta) = \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{1-x_i}.$$

   (Check that you understand why this formula is correct!) The *likelihood* of θ is the same quantity, only considered as a function of θ, with D fixed, i.e.

$$L(\theta \mid D) = P(D \mid \theta).$$

   (a) Calculate the *log likelihood* $\ell = \log L$, and express it in terms of M and N, where M is the number of heads observed, assuming $0 < M < N$.

   (b) Find the *maximum likelihood estimate* of θ given the observed data D, i.e. the value of θ that maximizes $L(\theta \mid D)$. *Hint: use calculus. Remember that maximizing L is equivalent to maximizing ℓ. (Make sure you understand why.) In this case (and often), it's easier to maximise ℓ.*

   Remember to cite any source you use if it answers this specific question or one very similar.

Question 3 on next page.

3. We reconsider the problem from the Week 4 Lab of classifying a penguin as male or female. You will need the dataset from that lab, which is from here: `https://raw.githubusercontent.com/mwaskom/seaborn-data/master/penguins.csv`. As before we consider only the Gentoo penguins. Unlike in the lab, we consider here only flipper length as a predictor of sex. [5pts]

   (a) Suppose that flipper lengths of males follow a Gaussian (i.e. Normal) distribution $\mathcal{N}(\mu_M, \sigma_M^2)$, for some unknown parameters μ_M, σ_M^2 (the mean and variance); and suppose that flipper lengths of females follow another Gaussian distribution $\mathcal{N}(\mu_F, \sigma_F^2)$. (These assumptions comprise a *model family*, i.e. an *unfitted model*.) Estimate all of these parameters: $\mu_M, \sigma_M^2, \mu_F, \sigma_F^2$ from data, using any standard method and/or any standard python package. We now have a *model*, i.e. a *fitted model*, i.e. the parameters have been *fitted* to the data, i.e. *learned* from the data. The model consists of two conditional probability densities,

$$p\left(\text{flipper length} \mid \text{sex} == \text{male}\right),$$
$$p\left(\text{flipper length} \mid \text{sex} == \text{female}\right).$$

   (b) Suppose that Gentoo penguins are known to be approximately 50% male and 50% female. This is our *prior* belief about the sex of any new penguin we consider, before we observe anything about them. We observe a new penguin, "Bela" and measure their flipper length to be 240mm. Using your model from above, and a function such as `scipy.stats.norm.pdf`, compute

$$p\left(\text{flipper length} == 240 \mid \text{sex} == \text{male}\right),$$
$$p\left(\text{flipper length} == 240 \mid \text{sex} == \text{female}\right).$$

   Use Bayes' rule to compute your posterior belief about Bela's sex. (If you prefer, you can do part (c) first, and use that to answer this part (b).) *Technical note: Bayes' theorem works for probability densities too.*

   (c) Write a function `bayes_classify()` that automates the calculation in part (b). It should take as inputs: a flipper length (in mm), and the parameters that you fitted in part (a); and give as output a vector of length two containing: the posterior probability that the penguin is female, and the posterior probability that the penguin is male. This is an example of a *Bayes classifier*, which is a kind of *probabilistic classifier* because it returns posterior probabilities and not just a single deterministic prediction.

   (d) Apply your classifier to all of the Gentoo penguins in the dataset and produce a chart showing flipper length, actual sex and predicted sex.

   Note that we are evaluating a classifier on its training set, which is usually not as informative as using a held-out test set; see the last part of the Week 4 Lab, and Assignment 4.