Example sheet 1

Learning with probability models
Foundations of Data Science—DJW—2019/2020

Question 1. Sketch the cumulative distribution function, and calculate the density function, for this continuous random variable:

```
def rx():
    u = random.random()
    return u * (1-u)
```

[Hint. See Exercise 3.3, from lecture 2. Sketch a graph of $u(1-u)$ as a function of u. For what ranges of u is $u(1-u) \leq y$? What is the probability that the random variable $U \sim U[0,1]$ lies in these ranges?]

Question 2. We wish to implement a random variable whose cumulative distribution function $F(x) = \mathbb{P}(X \leq x)$ is given by the function below. Here, a and b are parameters in the range $[0,1]$. Sketch $F(x)$, and give code to generate such a random variable.

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ bx/a & \text{if } 0 \leq x \leq a \\ b + (1-b)(x-a)/(1-a) & \text{if } a < x \leq 1 \\ 1 & \text{if } x > 1. \end{cases}$$

[Hint. See slide 10 from lecture 2. Also see the solution to mock exam question 1(b), in a video posted on Moodle, which suggests inventing a "mixture of uniforms". Try answering the question first for parameters $a = 1/2$, $b = 1/4$, and after that go on to the general case.]

Question 3. Given a dataset $(x_1, \ldots, x_n)$, we wish to fit a Poisson distribution. This is a discrete random variable with a single parameter $\lambda > 0$, called the rate, and

$$\Pr(x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \quad \text{for } x \in \{0, 1, 2, \ldots\}.$$

Show that the maximum likelihood estimator for λ is $\hat{\lambda} = n^{-1} \sum_{i=1}^n x_i$. *[Hint. This is a question about learning generative models. See section 1.5 exercise 1.7.]*

Question 4. Given a dataset [3,2,8,1,5,0,8], we wish to fit a Poisson distribution. Give code to achieve this fit, using `scipy.optimize.fmin`. *[Hint. See section 1.2 exercise 1.4. Also, if you use numpy, watch out for which variables in your code are vectors and which are scalars.]*

Question 5. Given a dataset $(x_1, \ldots, x_n)$, we wish to fit the Uniform$[0, \theta]$ distribution, where θ is unknown. By writing the density with explicit boundaries,

$$\Pr(x \mid \theta) = \frac{1}{\theta} 1_{x \geq 0} 1_{x \leq \theta} \quad \text{for } x \in \mathbb{R},$$

show that the maximum likelihood estimator is $\hat{\theta} = \max_i x_i$.

[Hint. In any question where the range of the random variable depends on unknown parameters, it's a good idea to include the boundaries explicitly in your density function, using an indicator function. See lecture 2 slides 10–11. A neat thing about indicator functions is that

$$1_{\xi \geq a} \times 1_{\xi \geq b} = 1_{\xi \geq a \text{ and } \xi \geq b} = 1_{\xi \geq \max(a,b)}.$$

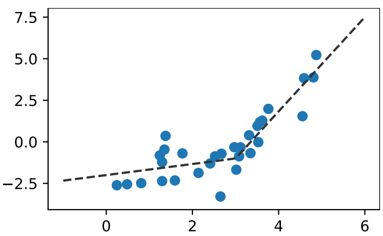
Use indicator functions, including this equation, in your answer. It will help make sure you're not missing any corner cases.]

Question 6 (A/B testing). Your company has two systems which it wishes to compare, A and B. It has asked you to compare the two, on the basis of performance measurements $(x_1, \dots, x_m)$ from system A and $(y_1, \dots, y_n)$ from system B. Any fool using Excel can just compare the averages, $\bar{x} = m^{-1} \sum_{i=1}^{m} x_i$ and $\bar{y} = n^{-1} \sum_{i=1}^{n} y_i$, but you are cleverer than that and you will harness the power of Machine Learning.

Suppose the x_i are drawn from $X \sim \text{Normal}(\mu, \sigma^2)$, and the y_i are drawn from $Y \sim \text{Normal}(\mu + \delta, \sigma^2)$, and all the samples are independent, and μ, δ, and σ are unknown. Find maximum likelihood estimators for the three unknown parameters. *[Hint. See exercise 1.8. When you do maximum likelihood estimation, you are optimizing* $\log \text{lik}(\textit{params} | \textit{data})$*, and the data should include absolutely all data that can shed light on the params. Don't estimate σ from the x_i alone—you should find a way to estimate it from both the x_i and the y_i, since both of them are informative about it.]*

Question 7. Let x_i be the population of city i, and let y_i be the number of crimes reported. Consider the model $Y_i \sim \text{Poisson}(\lambda x_i)$, where $\lambda > 0$ is an unknown parameter. Find the maximum likelihood estimator $\hat{\lambda}$. *[Hint. This is a question about supervised learning. See section 1.6 exercise 1.11.]*

Question 8. We wish to fit a piecewise linear line to a dataset, as shown below. The inflection point is given, and we wish to estimate the slopes and intercepts. Explain how to achieve this using a linear modelling approach.



Hint. See sections 2.2.1 and 2.2.2. Your model should represent a continuous line with an inflection point, not two separate lines. As a sanity check, you could implement your model equation and plot it. The code below illustrates a model that fails the sanity check:

```
def pred(x, m1,c1,m2,c2, inflection_x=3):
    e = numpy.where(x <= inflection_x, 1, 0)
    return e*(m1*x + c1) + (1-e)*(m2*x+c2)
x = numpy.linspace(0,5,1000)
plt.plot(x, pred(x, m1=0.5,c1=0,m2=1,c2=2))
```

Question 9. For the climate data from section 2.2.5 of lecture notes, we proposed the model

$$\mathsf{temp} \approx \alpha + \beta_1 \sin(2\pi \mathsf{t}) + \beta_2 \cos(2\pi \mathsf{t}) + \gamma \mathsf{t}$$

in which the $+\gamma\mathsf{t}$ term asserts that temperatures are increasing at a constant rate. We might suspect though that temperatures are increasing non-linearly. To test this, we can create a non-numerical feature out of t by

$$\mathsf{u} = \text{'decade\_'} + \mathsf{str}(\mathsf{math.floor}(\mathsf{t}/10)) + \text{'0s'}$$

(which gives us values like 'decade_1980s', 'decade_1990s', etc.) and fit the model

$$\mathsf{temp} \approx \alpha + \beta_1 \sin(2\pi \mathsf{t}) + \beta_2 \cos(2\pi \mathsf{t}) + \gamma_{\mathsf{u}}.$$

Write this as a linear model, and give code to fit it. *[See section 2.2.2. You should explain what the feature vectors are, then give a one-line command to estimate the parameters.]*

Question 10. I have two feature vectors

$$\mathsf{gender} = [f, f, f, f, m, m, m], \qquad \mathsf{eth} = [a, a, b, w, a, b, b]$$

and I one-hot encode them as

$$g_1 = [1,1,1,1,0,0,0] \qquad\qquad e_1 = [1,1,0,0,1,0,0]$$
$$g_2 = [0,0,0,0,1,1,1] \qquad\qquad e_2 = [0,0,1,0,0,1,1]$$
$$e_3 = [0,0,0,1,0,0,0]$$

Are these five vectors $\{g_1, g_2, e_1, e_2, e_3\}$ linearly independent? If not, find a linearly independent set of vectors that spans the same feature space. *[Hint. See section 2.5 exercise 2.3.]*

Question 11. For the police stop-and-search dataset in section 2.5 example 2.4, we wish to investigate intersectionality in police bias. We propose the linear model

$$1[\mathsf{outcome{=}find}] \approx \alpha_{\mathsf{gender}} + \beta_{\mathsf{eth}}.$$

Write this as a linear model using one-hot coding. Are the parameters identifiable? If not, rewrite the model so they are, and interpret the parameters of your model. *[Hint. Section 2.5 example 2.4.]*

Supplementary question sheet 1

Learning with probability models
Foundations of Data Science—DJW—2019/2020

These questions are not intended for supervision (unless your supervisor directs you otherwise). Some of them are longer form exam-style questions, which you can use for revision. Others, labelled $*$, ask you to think outside the box.

Question 12 (Cardinality estimation).

(a) Let T be the maximum of m independent Uniform$[0,1]$ random variables. Show that $\mathbb{P}(T \leq t) = t^m$. Find the density function $\mathrm{Pr}_T(t)$. *Hint. For two independent random variables U and V,*

$$\mathbb{P}\bigl(\max(U,V) \leq x\bigr) = \mathbb{P}(U \leq x \text{ and } V \leq x) = \mathbb{P}(U \leq x)\,\mathbb{P}(V \leq x).$$

(b) A common task in data processing is counting the number of unique items in a collection. When the collection is too large to hold in memory, we may wish to use fast approximation methods, such as the following: Given a collection of items $a_1, a_2, \ldots$, compute the hash of each item $x_1 = h(a_1), x_2 = h(a_2), \ldots$, then compute $t = \max_i x_i$.

If the hash function is well designed, then each x_i can be treated as if it were sampled from Uniform$[0,1]$, and unequal items will yield independent samples..

The more unique items there are, the larger we expect t to be. Given an observed value t, find the maximum likelihood estimator for the number of unique items. *[Hint. This is about finding the mle from a single observation, as in exercise 1.1.]*

`http://blog.notdot.net/2012/09/Dam-Cool-Algorithms-Cardinality-Estimation`

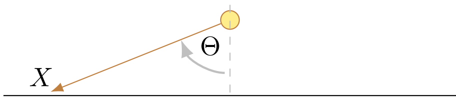
Question 13*. Sketch the cumulative distribution functions for these two random variables. Are they discrete or continuous?

```
def rx():
    u = random.random()
    return 1/u
def ry():
    u2 = random.random()
    return rx() + math.floor(u2)
```

[Hint. For intuition, use simulation. Generate say 10,000 samples, and plot a histogram, then a plot of "how many are $\leq x$" as a function of x.]

Question 14. A point lightsource at coordinates $(0,1)$ sends out a ray of light at an angle Θ chosen uniformly in $[-\pi/2, \pi/2]$. Let X be the point where the ray intersects the horizontal line through the origin. What is the density of X? *[Hint. See exercise 3.3, from lecture 2.]*

Note: This random variable is known as the Cauchy distribution. It is unusual in that it has no mean.



Question 15. As an alternative to the model from question 9, we might suspect that temperatures are increasing linearly up to 1980, and that they are increasing linearly at a different rate from 1980 onwards. Devise a linear model to express this, using your answer to question 8, and fit it. Plot your fit. *[Hint. Sample code for plotting a fit is shown in section 2.2.4.]*

Question 16 (A/B testing). The dataset for question 6 has been presented to you as a spreadsheet with $m+n$ rows and two columns, one column `measurement` containing $(x_1, \ldots, x_m, y_1, \ldots, y_n)$, and another column system whose entries are either A or B indicating which system the measurement came from.

Write the probabilistic model from question 6 as a linear model, with coefficients μ and δ. Explain what your feature vectors are. *[Hint. Use the approach of section 2.4.]*

Question 17. Here are two different models for the climate data:

$$\mathsf{temp} \approx \alpha + \beta_1 \sin(2\pi \mathsf{t}) + \beta_2 \cos(2\pi \mathsf{t}) + \gamma \mathsf{t}$$

and

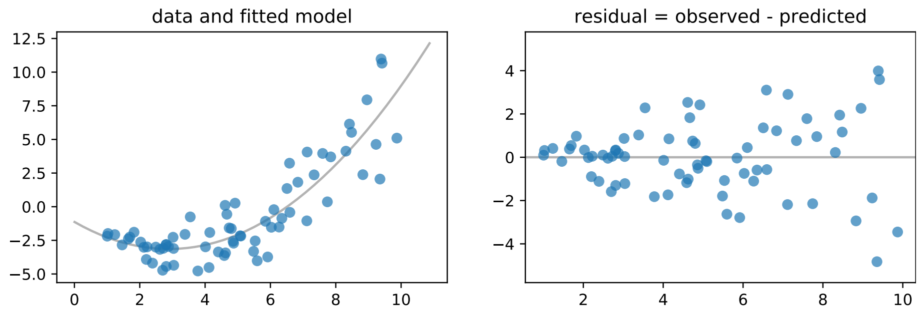
$$\mathsf{temp} \approx \alpha + \beta_1 \sin(2\pi \mathsf{t}) + \beta_2 \cos(2\pi \mathsf{t}) + \gamma(\mathsf{t} - 2000).$$

The first model produces a fitted value $\alpha = -63.9$°C and the second model produces a fitted value $\alpha = 10.5$°C. Why the difference? Which is correct? *[The answer is in section 2.2.5. But try to answer yourself, before looking it up.]*

Question 18 (Heteroscedasticity). We are given a dataset[1] with predictor x and label y and we fit the linear model

$$y_i \approx \alpha + \beta x_i + \gamma x_i^2.$$

After fitting the model using the least squares estimation, we plot the residuals $\varepsilon_i = y_i - (\hat{\alpha} + \hat{\beta} x_i + \hat{\gamma} x_i^2)$.



(a) Describe what you would expect to see in the residual plot, if the assumptions behind linear regression are correct.

(b) This residual plot suggests that perhaps $\varepsilon_i \sim \text{Normal}\big(0, (\sigma x_i)^2\big)$ where σ is an unknown parameter. Assuming this is the case, give pseudocode to find the maximum likelihood estimators for α, β, and γ.

[Hint. This question is asking you to reason about a custom probability model, in the style of section 2.4. A model with unequal variances is called 'heteroscedastic'.]

Question 19. Let $(F_1, F_2, F_3, \dots) = (1, 1, 2, 3, \dots)$ be the Fibonacci numbers, $F_n = F_{n-1} + F_{n-2}$. Define the vectors f, f_1, f_2, and f_3 by

$$\begin{aligned} f &= [F_4, F_5, F_6, \dots, F_{m+3}] \\ f_1 &= [F_3, F_4, F_5, \dots, F_{m+2}] \\ f_2 &= [F_2, F_3, F_4, \dots, F_{m+1}] \\ f_3 &= [F_1, F_2, F_3, \dots, F_m] \end{aligned}$$

for some large value of m. If you were to fit the linear model

$$f \approx \alpha + \beta_1 f_1 + \beta_2 f_2$$

what parameters would you expect? What about the linear model

$$f \approx \alpha + \beta_1 f_1 + \beta_2 f_2 + \beta_3 f_3?$$

[Hint. Are the feature vectors linearly independent?]

[1] https://teachingfiles.blob.core.windows.net/datasets/heteroscedasticity.csv