Example sheet 1

Learning with probability models

Data Science—DJW—2021/2022

Question 1. 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

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

Show that the maximum likelihood estimator for λ is $\hat{\lambda} = n^{-1} \sum_{i=1}^n x_i$.

Question 2. 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`.

Question 3. The dataset in question 2 comes from counts of radioactive particle emissions. The technician calls you and says that the counter's display is defective and that any value larger than 20 just displays as a 20. You therefore decide to model the datapoints as a truncated Poisson distribution,

$$\Pr(x\,;\,\lambda) = \begin{cases} \lambda^x e^{-\lambda}/x! & \text{if } x \in \{0, 1, \ldots, 19\} \\ 1 - \sum_{r=0}^{19} \lambda^r e^{-\lambda}/r! & \text{if } x = 20 \\ 0 & \text{if } x > 20 \end{cases}$$

(a) Your engineer friend thinks one should use unbiased estimators rather than maximum likelihood estimators. Show that for the Poisson probability model in question 1 the maximum likelihood estimator $\hat{\lambda} = n^{-1} \sum_{i=1}^n x_i$ is unbiased.

(b) Explain why, for the dataset in question 2, $\hat{\lambda} = n^{-1} \sum_{i=1}^n x_i$ is the maximum likelihood estimate for the truncated Poisson model.

(c) Show that $\hat{\lambda} = n^{-1} \sum_{i=1}^n x_i$ is *not* an unbiased estimator in the truncated Poisson model.

(d) Which do you think one should use, maximum likelihood estimators or unbiased estimators? Why?

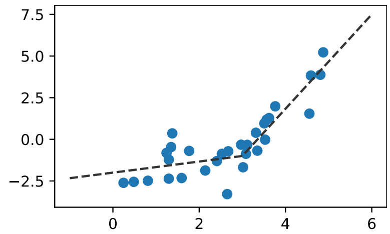
Question 4. Given a dataset $(x_1, \ldots, x_n)$, we wish to fit the Uniform$[0, \theta]$ distribution, where θ is unknown. Show that the maximum likelihood estimator is $\hat{\theta} = \max_i x_i$.

Question 5 (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 $(x_1, \ldots, x_m)$ from system A and $(y_1, \ldots, y_n)$ 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.

Question 6. 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 $\hat{\lambda}$.

Question 7. 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.



*Note. As a sanity check, you should implement your model formula as a function and plot it. Here's a function that **fails** the 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 8. For the climate data from section 2.2.5 of lecture notes, we proposed the model

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

in which the $+\gamma\texttt{t}$ 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

```
u = 'decade_' + str(math.floor(t/10)) + '0s'
```

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

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

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

Question 9. I have two feature vectors

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

and I one-hot encode them as

$$\begin{aligned} g_1 &= [1,1,1,1,0,0,0] & e_1 &= [1,1,0,0,1,0,0] \\ g_2 &= [0,0,0,0,1,1,1] & e_2 &= [0,0,1,0,0,1,1] \\ & & e_3 &= [0,0,0,1,0,0,0] \end{aligned}$$

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.

Question 10. For the police stop-and-search dataset in section 2.6, we wish to investigate intersectionality in police bias. We propose the linear model

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

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.

[Optional.] Fit the model and report your findings. Code to read the data and prepare `eth` and `gender` features can be found at `https://github.com/damonjw/datasci/blob/master/stop-and-search.ipynb`.

Hints and comments

Question 1. This is a question about learning generative models. See section 1.6.

Question 2. See section 1.4. What parameter transform is needed here? Also, if you use numpy, watch out for which variables in your numpy code are vectors and which are scalars.

Question 3. Unbiased estimators were taught in IA Intoduction to Probability. Look up the expected value of $X \sim \text{Poisson}(\lambda)$ on Wikipedia.

For part (b), *don't* try to derive the maximum likelihood estimator for an *arbitrary dataset* $(x_1, \dots, x_n)$. Instead just write out the likelihood function for *this particular dataset*, and ask yourself how your formula relates to question 1.

For part (c), you can answer the question numerically: write a function that computes $\mathbb{E}\,X = \sum_x x \Pr(x)$ when X has the truncated distribution, and show that there is at least one λ for which $\mathbb{E}\,X \neq \lambda$. This is enough to prove that the estimator isn't unbiased. (Alternatively, there is a slick piece of algebra that works for any λ.)

Question 4. This is a question about generative models, section 1.6. You will also need to use the indicator function trick, from section 1.3 exercise 1.3.6.

Question 5. You can treat this as a pure example of maximum likelihood estimation, as described in section 1.3. You are maximizing $\Pr(\text{data}\,;\,\text{params})$, where 'data' should include absolutely all data given to you in the question. The data here is $(x_1, \dots, x_m, y_1, \dots, y_n)$, and the params are (μ, δ, σ). Don't try to estimate μ from the x_i alone.

You can also treat this along the lines of section 2.4, as a linear model with a probabilistic interpretation. See the discussion in section 2.2 about designing features to compare groups.

Question 6. Supervised learning. See section 1.8 and the example of binomial regression.

Question 7. See section 2.2 on feature design, and the example of a step function.

Question 8. See section 2.2 on feature design, and the subsection on step function responses.

Question 9. See section 2.5 and the exercise about finding a linearly independent subset.

Supplementary question sheet 1

Learning with probability models
Data Science—DJW—2021/2022

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. Some others, labelled ∗, ask you to think outside the box.

Question 11 (Numerical optimization). Fit the model

$$\texttt{Petal.Length} \approx \alpha - \beta(\texttt{Sepal.Length})^\gamma, \qquad \gamma > 0$$

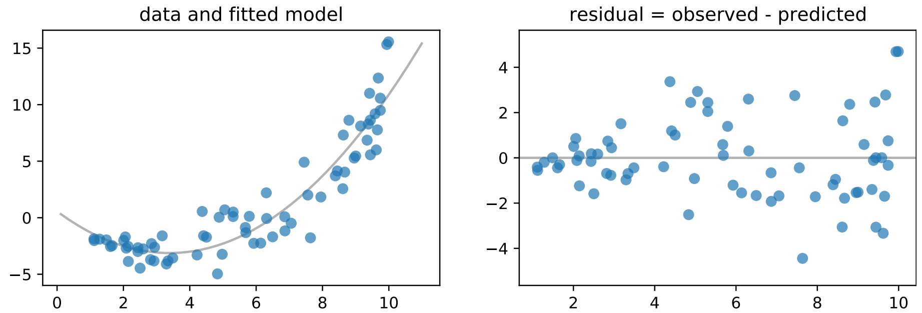
by minimizing the mean square error. *[Hint. This isn't a linear model, so just use `scipy.optimize.fmin`.]*

Question 12. As an alternative to the model from question 8, 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 7, and fit it. Plot your fit. *[Hint. Sample code for plotting a fit is shown in section 2.1.]*

Question 13 (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 14. 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

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

[1] `https://www.cl.cam.ac.uk/teaching/2122/DataSci/data/heteroscedasticity.csv`

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?]

Question 15*. For the police stop-and-search data from section 2.6, consider the model

$$1[\texttt{outcome} = \text{find}] = \alpha + \sum_{k \neq \text{White}} \beta_k \big(1[\texttt{eth} = k] - 1[\texttt{eth} = \text{White}]\big).$$

Interpret the parameters. *[Hint. What is the predicted value for each ethnicity? What is the average prediction across all ethnicities?]*

Question 16*. 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 17*. Is it possible for a continuous random variable to have a probability density function that approaches ∞ at some point in the support? Is it possible to have this and also have finite mean and variance?