Example sheet 1

Learning with probability models
Data Science—DJW—2025/2026

Some of the questions ask for pseudocode. I suggest you implement your code, and test it using the online tester. There is a notebook with templates for answers and instructions for submission on the course materials webpage.

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. Give pseudocode to fit the model of question 1, using `scipy.optimize.fmin`. *[Optional.]* If you want to test your code using the online tester, fill in the answer template for `PoissonModel`.

In practice it'd be daft to use numerical optimization when we have an exact formula for the answer. But it's good to get used to numerical optimization, and it's good to test your code on a problem where you what the answer should be.

Question 3. 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 4 (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.

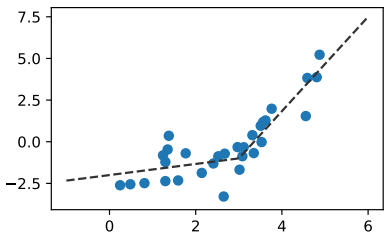
Question 5. Given a dataset $[x_1, \ldots, x_n]$, we wish to estimate the signal-to-noise ratio $\lambda = \mu/\sigma$ from a Normal(μ, σ^2) model. Your engineer friend proposes rewriting the model as Normal$(\lambda\sigma, \sigma^2)$ and estimating λ from this. Give a formula for the maximum likelihood estimator $\hat{\lambda}$, and explain why your friend is silly.

Question 6. Let x_i be the population of city $i \in \{1, \ldots, n\}$, 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 (Likelihood notation). Let $X \sim \text{Bin}(4, 1/2)$. What is $\Pr_X(X)$?

Question 8. We are given a dataset of (x_i, y_i) pairs and we wish to model y as a continuous function of x, using a piecewise linear response function as shown below. The response function consists of two straight lines that meet at an inflection point; the x-coordinate of the inflection point is given. Explain how to achieve this using a linear modelling approach.

[Optional.] If you want to test your code using the online tester, fill in the answer template for `PiecewiseLinearModel`.



Question 9. 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.]*

[Optional.] If you want to test your code using the online tester, fill in the answer template for `StepPeriodicModel`.

Question 10. 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

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

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.

Question 11. 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.] If you want to explore this dataset yourself, there is code on the course materials webpage to download the data from `data.police.uk` and to prepare the `eth` and `gender` features.

Hints and comments

Question 1. This is a question about fitting parameters from a dataset, as in section 1.3. Formally it's unsupervised learning. There are more examples in section 1.7.

Question 2. See section 1.4. What parameter transform is needed here, to perform a maximization over the restricted domain $\lambda > 0$? How many maxima does the likelihood function have, and how might you choose a sensible starting point for the numerical optimization to make sure it finds a global maximum? Also, if you use numpy, watch out for which variables in your numpy code are vectors and which are scalars.

Question 3. This is another question about maximum likelihood estimation on datasets, also known as unsupervised learning, like question 1. You will also need to use the indicator function trick, from section 1.3 exercise 1.3.6.

Question 4. You can treat this as a pure example of maximum likelihood estimation, as described in section 1.3. You are maximizing Pr(data ; params). The 'data' should include absolutely all data given to you in the question, as in exercise 1.3.8. The data here for this question is $(x_1, \ldots, x_m, y_1, \ldots, y_n)$, 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 5. Look at plug-in estimation, exercises 1.3.7 and 1.3.8.

Question 6. Supervised learning. See section 1.7 and the example of binomial regression.

Question 7. This is a puzzle to see if you've gotten your head around the likelihood notation in section 1.5. You need to think step by step through the logic of the notation, rather than jumping to conclusions …

(a) What is $\Pr_X$? It's a function. Write out exactly what the function is, listing all possible inputs and the corresponding outputs.

(b) What is X? It's a random variable. Write out all the possible values it can take, and their probabilities.

(c) What do you get if you apply a function to a random variable? You get another random variable. Therefore $\Pr_X(X)$ must be a random variable. Simply list the possible values it can take, and their probabilities.

Question 8. See section 2.2 on feature design.

If you simply fit two separate straight lines (i.e. you estimate four parameters, a slope and an intercept for each side of the inflection point), then your two straight lines might not meet. You will need to propose a model with THREE parameters, so that whatever values it fits for those three parameters you still end up with a continuous response function. Therefore you need to construct a linear model with THREE feature vectors, each of them weighted by one of your three parameters.

Question 9. See section 2.2 on feature design, and the subsection on step function responses.

Question 10. See section 2.5 and the exercise about finding a linearly independent subset.

Supplementary question sheet 1

There supplementary 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 (Maximum likelihood estimation versus unbiasedness). There is a dataset $[x_1, \ldots, x_n]$ of non-negative numbers. Two colleagues disagree about how to model it. Prof Chalk thinks we should model them as independent Uniform$[0, \theta]$ random variables, whereas Dr Cheese prefers Uniform$[0, \lambda]$.

- Find maximum likelihood estimators for θ and λ, call them $\hat{\theta}$ and $\hat{\lambda}$. What is the relationship between $\hat{\theta}$ and $\hat{\lambda}$?
- For $X_1, \ldots, X_n \sim$ Uniform$[0, \theta]$, find the cdf of $\max_i X_i$, and the pdf, and hence find $\mathbb{E} \max_i X_i$.
- Find unbiased estimators for θ and λ, call them $\bar{\theta}$ and $\bar{\lambda}$. What is the relationship between $\bar{\theta}$ and $\bar{\lambda}$?

Question 13 (Maximum likelihood estimation versus unbiasedness)*. Suppose you've fitted a Poisson model as in question 1 to the dataset $[3, 2, 8, 1, 5, 0, 8]$, which comes from counts of radioactive particle emissions. The next day, a 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 use a different model: you model the datapoints as coming from 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.[1]

(b) Explain why, for the concrete dataset specified above, $\hat{\lambda} = n^{-1} \sum_{i=1}^n x_i$ is the maximum likelihood estimate for the truncated Poisson model. *[Hint. Don't try to derive the maximum likelihood estimator for an* arbitrary dataset $[x_1, \ldots, x_n]$. *Instead just write out the likelihood function for this particular dataset, and ask yourself how your formula relates to question 1.]*

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

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

Question 14 (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`*.]*

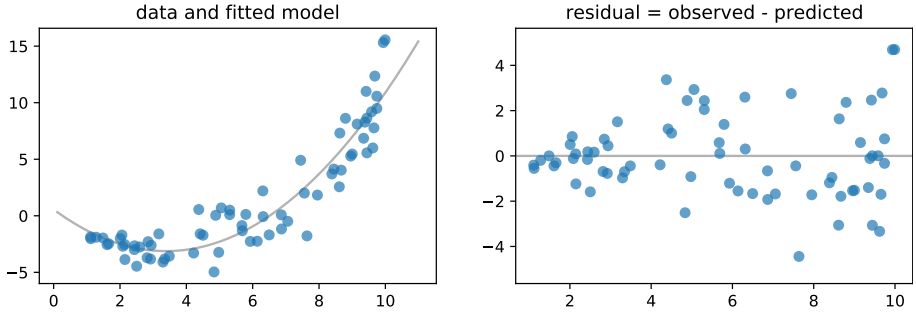
Question 15. As an alternative to the model from question 9, we might suspect that temperatures are increasing linearly up to 1980, and that they are increasing linearly at a different rate from 1980 onwards. Devise a linear model to express this, using your answer to question 8, and fit it. Plot your fit.

[1] What does it mean for $\hat{\lambda}$ to be an unbiased estimator? It means that if you apply the estimator to a random dataset $X_1, \ldots, X_n$, where each item in the dataset is drawn from Poisson(λ), then $\mathbb{E}\,\hat{\lambda} = \lambda$.

Question 16 (Heteroscedasticity). We are given a dataset[2] 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 17. 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?]

Question 18*. 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 19*. Sketch the cumulative distribution functions for these two random variables. Are they discrete or continuous?

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

```
def rx():
    u = random.random()
    return 1/u
def ry():
    x = rx()
    i = random.random() < 0.5
    return x if i else math.floor(x)
```

[Hint. For intuition, use simulation. Rewrite the code in numpy vectorized style. Then generate say 10,000 samples, and plot a histogram, then a plot of "how many are $\leq x$" as a function of x.]

Question 20*. 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?