Exercise Sheet 6: Data Science Methods

Technische Universität München
Winter semester 2024-2025

Instructions

You can submit your solutions via email to `alan.zander@tum.de` by 13/02/25.

1 Exercise: Numerical Integration

Write a function in Python that implements the trapezoid method in one dimension for estimating the integral of a function f in a given interval (a, b). For a given subinterval $[x_i, x_{i+1}]$, we can approximate the area under f by

$$\int_{x_i}^{x_{i+1}} f(x)\mathrm{d}x \approx \frac{f(x_i) + f(x_{i+1})}{2}\Delta x,$$

where $\Delta x \equiv x_{i+1} - x_i$. We can then sum over all subintervals to obtain $\int_a^b f(x)\mathrm{d}x$.

2 Exercise: Model Selection

The Bayes factor K is a measure used in Bayesian statistics to compare two competing models, M_0 and M_1. It is defined as the ratio of the marginal likelihoods (a.k.a. the model-specific evidence) of the two models:

$$K = \frac{P(\text{Data} \mid M_0)}{P(\text{Data} \mid M_1)}.$$

The marginal likelihood for a model is the probability of the observed data under that model, averaged over all unknown parameters. For a model M, it is given by:

$$P(\text{Data} \mid M) = \int P(\text{Data} \mid \theta, M) P(\theta \mid M)\, d\theta,$$

where $P(\text{Data} \mid \theta, M)$ is the likelihood of the data given the parameters θ, and $P(\theta \mid M)$ is the prior distribution of the parameters under model M.

Task

Consider the ratio between the Euro and the US Dollar (EUR/USD) exchange rate. We aim to investigate two models that describe the probability of continuing a trend in daily returns. A ”green day” is defined as a day with a positive return, i.e. the exchange rate at the end of the day is higher than the exchange rate on the day before. A ”red day” is defined as a day with a negative return.

The two models under consideration are:

- **Model** M_0: The probability q of continuing the trend (a green day followed by another green day, or a red day followed by another red day) is fixed at $q = 0.5$.
- **Model** M_1: The probability q of continuing the trend is unknown and follows a uniform prior distribution over the interval $[0, 1]$.

Each day, we can look at the day's return, compare it to the return on the day before and determine whether or not they are both positive or negative. To collect some data, this Bernoulli experiment can be repeated every day for some period of time.

2.1 Load the file `EUR_USD_exchange_rate.csv` into your working environment to use it as the dataset for this analysis. Each row corresponds to the exchange rate at the end of one day. Determine the output of our Bernoulli experiment for each day (except for the first one, which has no day before).

2.2 Assuming that daily returns are independent of each other, calculate the Bayes factor K. For the integral over the success probability q use your implementation of the trapezoid method from Exercise 1. Is there a preference for one model over the other based on the data?

2.3 We want to compare this procedure to the classical approach. Construct the LRT and assuming Wilk's theorem applies (which it does), compute the p-value and decide whether or not the null hypothesis can be rejected with a significance level of $\alpha = 0.05$. Interpret the results.

3 Exercise: Markov Chain Monte Carlo (MCMC)

A random variable X is normally distributed $X \sim \mathcal{N}\left(|\theta|, \sigma^2\right)$ with known $\sigma^2 = 1$ and unknown θ. We want to determine the posterior distribution $P\left(\theta|\text{data}\right)$ via MCMC given some data.

3.1 Generate the sample that we will use as data by writing the following code:

```
# Generate data
import scipy as sp # Erase this line if already imported
theta0 = -3.3 # True value, however for us unknown
sigma = 1
n_total = 1000
data = sp.stats.norm.rvs(abs(theta0), sigma, size=n_total)
```

and illustrate the generated data.

3.2 Apply the Metropolis algorithm to estimate the posterior distribution $P\left(\theta|\text{data}\right)$. Assume a flat prior (uniform distribution $P(\theta) \sim U(-30, 30)$) and a Gaussian proposal distribution $g(y|x) \sim \mathcal{N}\left(x, \tau^2 = 5^2\right)$, where x represents the current state and y the proposed new state in the Markov chain.
Hint: Use the log-likelihood instead of the likelihood to avoid numerical underflow, and think about how the Metropolis acceptance ratio changes when working in log space.

3.3 Now try $\tau = 1$ and discuss your results in contrast to $\tau = 5$.