Exercise Set 2

Please write the name(s) of the student(s) to the answer sheet!

Problem 1

Learning objective: discrete distributions, parameter estimates, bootstrap. [5 points]

Background story

Mulvaney et al. (2020) tested 3330 persons not previously diagnosed with COVID-19 in Santa Clara County, California, using serological test for COVID-19 antibodies and found 50 (1.5%) positive results. At the time of testing, there were 956 confirmed COVID-19 cases in the county. The county has a population of about 1.9 million.

The serological test used had been validated as follows:

* The test was applied to 371 persons known to be negative, resulting to 369 true negatives and 2 false positives (estimated specificity 0.9946).
* The test was applied to 197 persons known to be positive, resulting to 178 true positives and 19 false negatives (estimated sensitivity 0.9035).

The authors concluded that the unadjusted prevalence of antibodies in the county was 1.5% (95CI 1.11-1.97%). The authors actually obtained a still higher prevalence by taking into account the fact that the tested persons were not fully representative of the population. The actual number of infected persons therefore appears to be orders of magnitude larger than the number of confirmed cases.

Reference: Mulaney et al. (2020) COVID-19 Antibody Seroprevalence in Santa Clara County, California. medRxiv 2020.04.14.20062463. <https://doi.org/10.1101/2020.04.14.20062463>

The actual machine learning problem

In the lectures we computed means and used normal distributions to model both the likelihood $p(x\mid\mu)$ and priors $p(\mu)$. Lets move from real numbers to binary world.

Replace the real numbers by binary numbers, i.e., instead of n real numbers consider n binary numbers $x_i\in\{0,1\}$, $i\in\{1,\ldots,n\}$, sampled i.i.d. from [Bernoulli distribution](https://en.wikipedia.org/wiki/Bernoulli_distribution) $p(x\mid \theta)$ parametrised by $\theta\in[0,1]$ (probability of $x=1$ being θ and probability of $x=0$ being $1-\theta$). Further assume that θ obeys prior distribution given by the Beta distribution $p(\theta)\propto\theta^{\alpha-1}(1-\theta)^{\beta-1}$, where $\alpha>0$ and $\beta>0$ are prior parameters. Now, consider a case where we have drawn n binary numbers of which k turned out to be ones and $n-k$ zeros. I.e., we are actually talking about [Binomial distribution](https://en.wikipedia.org/wiki/Binomial_distribution) which you get when you sample n binary numbers from Bernoulli distribution.

Tasks

* Write down the formulas for likelihood and posterior as a function of n, k, α, β, and θ. You can ignore multiplicative factors that are constant in θ (why?).

```
* Derive the ML and MAP estimates $\hat\theta_{ML}$ and $\hat\theta_{MAP}$ for
$\theta$. How can you interpret $\alpha$ and $\beta$ that appear in the formula for
$\hat\theta_{MAP}$?
* Are the estimates $\hat\theta_{ML}$ and $\hat\theta_{MAP}$ unbiased and/or
consistent?
* Plug in suitable $n$ and $k$ and compute ML and MAP estimates for the specificity
and sensitivity for the serological test described above.
* Then compute bootstap 95CI for the specificity and sensitivity.
* Let's consider null hypothesis that the prevalence is zero, i.e., all of the
observed 50 positives are actually false positives. There are many ways to
construct a statistical test to try to reject this null hypothesis. One approach is
to use [Fisher's Exact test](https://en.wikipedia.org/wiki/Fisher%27s_exact_test).
Compute the p-value given by the Fisher's exact test. Can you rule out the null
hypothesis?
* Sample values of specificity (i.e., $\theta$) from its posterior distribution by
using Stan. Compute 95CI (credible interval).

Hint: Fisher's exact test works on 2x2 contingency tables. You can think that the
rows of the table correspond to the validation test with 317 participants and the
test protocol with 3300 participants, respectively, and the columns to negative and
positive test result.

### Bonus task

You can try this if you have time.

* Use Stan and a Bayesian model that has 3 parameters: specificity, sensitivity,
and prevalence. The two validation tests depend on specificity and sensitivity,
respectively, and the observed 50 positives can be sampled from the Binomial
distribution with a parameter that can be computed from sensitivity, specificity,
and prevalence. Show samples from the posterior distribution of the three
parameters, e.g., by using pair plots and compute and report the credible intervals
for all three parameters.
```

Problem 2

Learning objective: theory of bootstrap. [2 points]

Read Section 5.2 of James et al.

Task

- Compute probability that a given observation is part of a bootstrap sample.

Hint: James et al., Section 5.4, Exercise 2, page 197. $\lim_{n\to\infty}(1-1/n)^n = 1/e \approx 0.368$.

Problem 3

Learning objective: regression models, generalisation, validation techniques. [5 points]

Read Section 5.1 of James et al.

As you know, in supervised learning we try to find a function that produces a good estimate of the dependent variable. Consider a dataset collected from Hyytiälä where our objective is to estimate CO2 fluxes in and out from the forest, given some covariates. The resulting regressor could, e.g., be used to fill in the gaps in the

CO2 flux measurements. In this dataset, the attribute FCO2 gives the CO2 flux and the other variables are from other measurements that can be used as covariates in our regressor.

Lets simulate missing data by choosing n=100 variables in random (see the code below) by using which we are allowed to train the regressor (called the *training set*). The remaining data items are used (*test set*).

You have heard that he following very good regressors that are all luckily available via Scipy (or R):

- OLS linear regression
- Regression tree
- Random forest
- SVM

Tasks

- Find still one more regression model (e.g., from sklearn) and add it to your list.
- Train all of the regressors on the training set and report the mean squared errors (MSE) on both the training and test sets. Which of the models performs best on the training set and on the test set? What does this tell about complexity of the respective model families?
- Next, split the 100 items in the training set in random to *new training set* of 50 items and to a *validation set* of 50 items. Train all of the regressors on the new training set and report MSE on the new training set, validation set, and the test set.
- Repeat the above, but instead of even split of the training set into new training set and validation set use cross-validation. Report your results.
- Which of the four regressors is the best? How does MSE on the training data compare to the error on the test set? How does the error on the validation set compare to the error on the test set? Could you do something with these regressor (on this training set) to make them perform better?

If you use Python you can use the training / test set split given below. I also give below pointers to suitable learning functions that you may want to use.

In [76]:

```
import numpy as np
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.svm import SVR

np.random.seed(42)

def listdiff(a,b):
    s = set(b)
    return [x for x in a if x not in s]

co2 = pd.read_csv("CO2_exchange.csv",index_col=0)
co2.columns = list(map(lambda x: x.replace("HYY_META.",""),co2.columns)) # get rid c

n = 100
itr = np.random.choice(co2.shape[0],n)  # training set index
ite = listdiff(range(co2.shape[0]),itr) # test set index

X_tr = np.array(co2.iloc[itr].drop("FCO2",axis=1)) # training set
y_tr = co2["FCO2"][itr]                            # training set
X_te = np.array(co2.iloc[ite].drop("FCO2",axis=1)) # test set
y_te = co2["FCO2"][ite]                            # test set

mse = lambda f, x, y: np.mean((f.predict(x)-y)**2)

fit = LinearRegression().fit(X_tr,y_tr)
print("LinearRegression: MSE_tr = %g  MSE_te = %g" % (mse(fit,X_tr,y_tr),mse(fit,X_t
```

```
LinearRegression: MSE_tr = 4.42859  MSE_te = 9.01516
```