Machine Learning Exercise Sheet 3 (Probabilistic Inference)

tags: `IN2064 Machine Learning`

Group 100 (Yi-Hsiang Fang, Hung-Ju Wu, Yu-Ju Chiu)

Problem 6

$$\begin{aligned}
\frac{d}{d\theta}\theta^t(1-\theta)^h &= t\theta^{t-1}(1-\theta)^h + \theta^t h(1-\theta)^{h-1}\cdot(-1) \\
&= \theta^{t-1}(1-\theta)^{h-1}[t(1-\theta)-h\theta] \\
\frac{d^2}{d\theta^2}\theta^t(1-\theta)^h &= t[(t-1)\theta^{t-2}(1-\theta)^h + \theta^{t-1}h(1-\theta)^{h-1}\cdot(-1)] - \\
&\quad h[t\theta^{t-1}(1-\theta)^{h-1} + \theta^t(h-1)(1-\theta)^{h-2}\cdot(-1)] \\
&= t(t-1)\theta^{t-2}(1-\theta)^h - 2th\theta^{t-1}(1-\theta)^{h-1} + h(h-1)\theta^t(1-\theta)^{h-2} \\
log\theta^t(1-\theta)^h &= tlog\theta + hlog(1-\theta) \\
\frac{d}{d\theta}log\theta^t(1-\theta)^h &= \frac{t}{\theta} - \frac{h}{1-\theta} \\
\frac{d^2}{d\theta^2}log\theta^t(1-\theta)^h &= -\frac{t}{\theta^2} - \frac{h}{(1-\theta)^2}
\end{aligned}$$

Problem 7

1. Take the logarithm of $f(\theta)$
   let $\theta^\star$ be an arbitary local maximun of $g(\theta) = logf(\theta)$
   $\Rightarrow g(\theta^\star) \geq g(\theta)$
2. Take exponential on both sides we get
   $f(\theta^\star) = exp(g(\theta^\star)) \geq exp(g(\theta)) = f(\theta)$
   $\Rightarrow f(\theta^\star) \geq f(\theta)$
   We can conclude that taking logarithm of any function will remain itsmaximun or minimum at the same point. Besides, it could reduce computational effort.

Problem 8

Since the postereor is Beta(m+a, l+b) distribution, the expectation of the distribution is $\frac{m+a}{m+l+a+b}$

$\mathbb{E}[\theta|\mathbf{D}] = \frac{m+a}{m+l+a+b} = \frac{m}{m+l+a+b} + \frac{a}{m+l+a+b}$

Let $\frac{a+b}{m+l+a+b} = \lambda$
$\frac{m}{m+l+a+b} = \frac{m+l}{m+l+a+b} \cdot \frac{m}{m+l} = (1-\lambda) \cdot \frac{m}{m+l}$

which $\frac{m}{m+l}$ is the maximum likelihood estimate

$\frac{a}{m+l+a+b} = \frac{a+b}{m+l+a+b} \cdot \frac{a}{a+b} = \lambda \cdot \frac{a}{a+b}$

which $\frac{a}{a+b}$ is the prior mean value of θ

Since $\frac{a+b}{m+l+a+b} = (1-\lambda) \cdot \frac{m}{m+l} + \lambda \cdot \frac{a}{a+b}$, the posterior mean is between the prior distribution and the maximum likelihood solution

Problem 9

$$
\begin{aligned}
p(\lambda|x) &= \frac{p(x|\lambda)p(\lambda|a,b)}{p(x)} \\
&\propto p(x|\lambda)p(\lambda|a,b) \\
&\propto \frac{\lambda^x \exp(-\lambda)}{x!} \cdot \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda) \\
&\propto \lambda^{x+a-1} \exp(-(b+1)\lambda) \\
\lambda_{MAP} &= \arg\max_\lambda p(\lambda|x) \\
&= \arg\max_\lambda \lambda^{x+a-1} \exp(-(b+1)\lambda) \\
&= \arg\max_\lambda \log(\lambda^{x+a-1} \exp(-(b+1)\lambda)) \\
&= \arg\max_\lambda (x+a-1)\log\lambda - (b+1)\lambda
\end{aligned}
$$

$$solve \frac{d}{d\lambda}[(x+a-1)\log\lambda - (b+1)\lambda] = 0:$$

$$
\begin{aligned}
\frac{d}{d\lambda}[(x+a-1)\log\lambda - (b+1)\lambda] &= \frac{x+a-1}{\lambda} - (b+1) = 0 \\
\lambda &= \frac{x+a-1}{b+1} \\
\therefore \lambda_{MAP} &= \frac{x+a-1}{b+1}
\end{aligned}
$$

Problem 10

Programming Task: Probabilistic Inference

In [1]:
```
import numpy as np
import matplotlib.pyplot as plt

from scipy.special import loggamma
%matplotlib inline
```

Your task

This notebook contains code implementing the methods discussed in `Lecture 3: Probabilistic Inference`. Some functions in this notebook are incomplete. Your task is to fill in the missing code and run the entire notebook.

In the beginning of every function there is docstring which specifies the input and and expected output. Write your code in a way that adheres to it. You may only use plain python and anything that we imported for you above such as `numpy` functions (i.e. no scikit-learn classifiers).

Exporting the results to PDF

Once you complete the assignments, export the entire notebook as PDF and attach it to your homework solutions. The best way of doing that is

1. Run all the cells of the notebook (`Kernel -> Restart & Run All`)
2. Export/download the notebook as PDF (`File -> Download as -> PDF via LaTeX (.pdf)`)
3. Concatenate your solutions for other tasks with the output of Step 2. On Linux you can simply use `pdfunite` , there are similar tools for other platforms too. You can only upload a single PDF file to Moodle.

Make sure you are using `nbconvert` **Version 5.5 or later** by running `jupyter nbconvert --version` . Older versions clip lines that exceed page width, which makes your code harder to grade.

Simulating data

The following function simulates flipping a biased coin.

In [2]:
```
# This function is given, nothing to do here.
def simulate_data(num_samples, tails_proba):
    """Simulate a sequence of i.i.d. coin flips.

    Tails are denoted as 1 and heads are denoted as 0.

    Parameters
    ----------
    num_samples : int
        Number of samples to generate.
    tails_proba : float in range (0, 1)
        Probability of observing tails.

    Returns
    -------
    samples : array, shape (num_samples)
        Outcomes of simulated coin flips. Tails is 1 and heads is 0.
    """
    return np.random.choice([0, 1], size=(num_samples), p=[1 - tails_proba, ta
ils_proba])
```

In [3]:
```
np.random.seed(123)  # for reproducibility
num_samples = 20
tails_proba = 0.7
samples = simulate_data(num_samples, tails_proba)
print(samples)
```

```
[1 0 0 1 1 1 1 1 1 1 1 1 1 0 1 1 0 0 1 1]
```

Important: Numerical stability

When dealing with probabilities, we often encounter extremely small numbers. Because of limited floating point precision, directly manipulating such small numbers can lead to serious numerical issues, such as overflows and underflows. Therefore, we usually work in the **log-space**.

For example, if we want to multiply two tiny numbers a and b, we should compute $\exp(\log(a) + \log(b))$ instead of naively multiplying $a \cdot b$.

For this reason, we usually compute **log-probabilities** instead of **probabilities**. Virtually all machine learning libraries are dealing with log-probabilities instead of probabilities (e.g. [Tensorflow-probability (https://www.tensorflow.org/probability)](https://www.tensorflow.org/probability) or [Pyro (https://pyro.ai)](https://pyro.ai)).

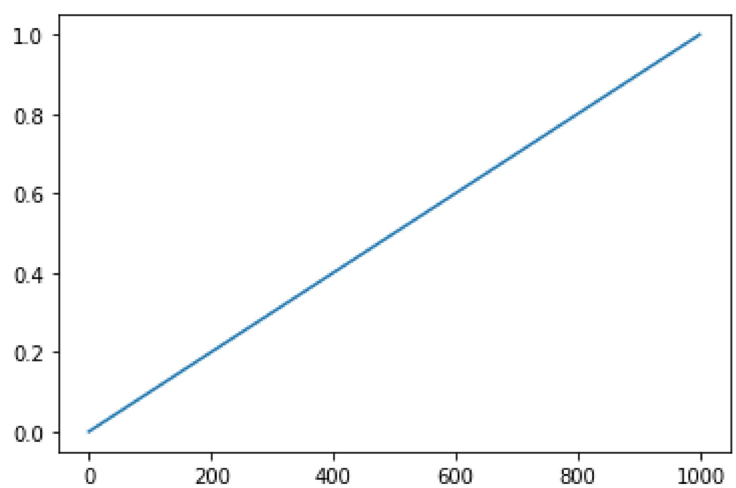
In [4]:
```
def helper(samples):
    num_tail = 0
    num_head = 0
    for sample in samples:
        if(sample == 1):
            num_tail += 1
        else:
            num_head += 1
    return num_tail, num_head
```

In [5]:
```
def compute_log_likelihood(theta, samples):
    """Compute log p(D | theta) for the given values of theta.

    Parameters
    ----------
    theta : array, shape (num_points)
        Values of theta for which it's necessary to evaluate the log-likelihoo
d.
    samples : array, shape (num_samples)
        Outcomes of simulated coin flips. Tails is 1 and heads is 0.

    Returns
    -------
    log_likelihood : array, shape (num_points)
        Values of log-likelihood for each value in theta.
    """
    ### YOUR CODE HERE ###

    # Count the number of tail and head
    num_tail, num_head = helper(samples)

    #print(num_tail/ (num_tail + num_head))
    log_likelihood = num_tail * np.log(theta) + num_head * np.log(1-theta)
    return log_likelihood
```

Task 1: Compute $\log p(\mathcal{D} \mid \theta)$ for different values of θ

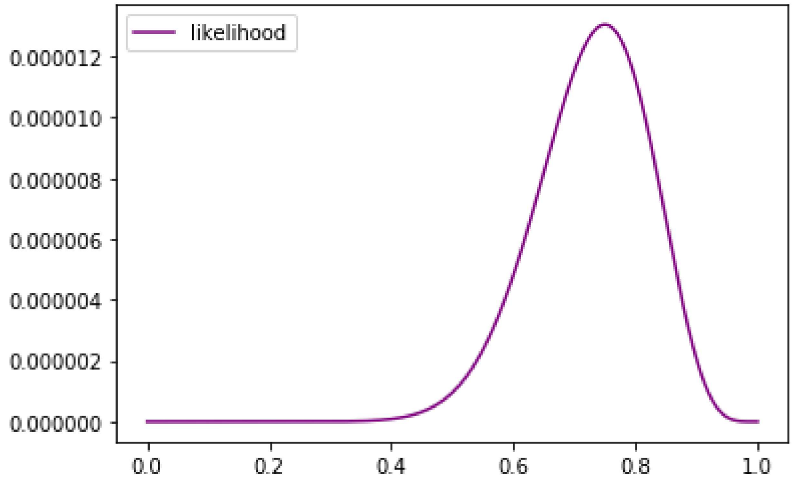
In [6]:
```
x = np.linspace(1e-5, 1-1e-5, 1000) # There are 1000 theta from probability 0.
00001 to 0.9999
plt.plot(x)                         # Then we can test which probabiltiy has t
he max likelihood to the samples
```

Out[6]: [<matplotlib.lines.Line2D at 0x1e3744c4b88>]



In [7]:
```
x = np.linspace(1e-5, 1-1e-5, 1000)
log_likelihood = compute_log_likelihood(x, samples)
likelihood = np.exp(log_likelihood)
plt.plot(x, likelihood, label='likelihood', c='purple')
plt.legend()
```

Out[7]: <matplotlib.legend.Legend at 0x1e3759600c8>



Note that the likelihood function doesn't define a probability distribution over θ --- the integral $\int_0^1 p(\mathcal{D} \mid \theta) d\theta$ is not equal to one.

To show this, we approximate $\int_0^1 p(\mathcal{D} \mid \theta) d\theta$ numerically using the rectangle rule (https://en.wikipedia.org/wiki/Riemann_sum).

In [8]:
```
# 1.0 is the length of the interval over which we are integrating p(D | theta)
int_likelihood = 1.0 * np.mean(likelihood)
print(f'Integral = {int_likelihood:.4}')
```

```
Integral = 3.068e-06
```

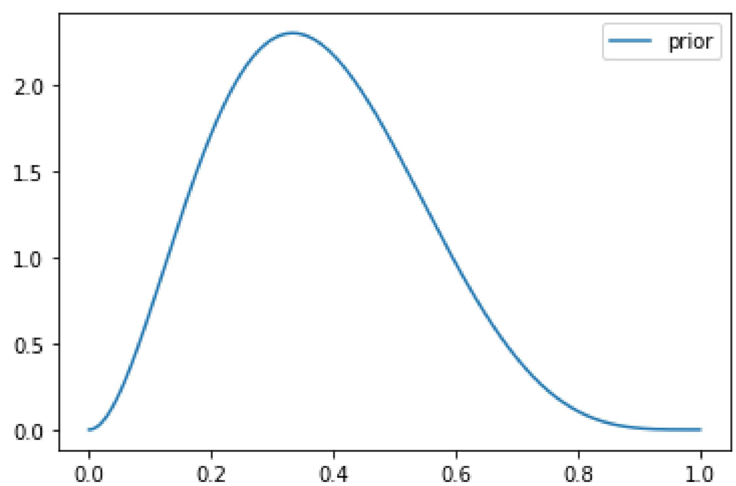
Task 2: Compute $\log p(\theta \mid a, b)$ for different values of θ

The function `loggamma` from the `scipy.special` package might be useful here. (It's already imported - see the first cell)

In [9]:
```
def compute_log_prior(theta, a, b):
    """Compute log p(theta | a, b) for the given values of theta.

    Parameters
    ----------
    theta : array, shape (num_points)
        Values of theta for which it's necessary to evaluate the log-prior.
    a, b: float
        Parameters of the prior Beta distribution.

    Returns
    -------
    log_prior : array, shape (num_points)
        Values of log-prior for each value in theta.

    """
    ### YOUR CODE HERE ###
    normalization_constant = loggamma(a+b) - loggamma(a) - loggamma(b)
    log_prior = normalization_constant + (a-1) * np.log(theta) + (b - 1) * np.
log(1 - theta)
    return log_prior
```

In [10]:
```
x = np.linspace(1e-5, 1-1e-5, 1000)
a, b = 3, 5

# Plot the prior distribution
log_prior = compute_log_prior(x, a, b)
prior = np.exp(log_prior)
plt.plot(x, prior, label='prior')
plt.legend()
```

Out[10]: <matplotlib.legend.Legend at 0x1e3759ef688>



Unlike the likelihood, the prior defines a probability distribution over θ and integrates to 1.

In [11]:
```
int_prior = 1.0 * np.mean(prior)
print(f'Integral = {int_prior:.4}')
```

```
Integral = 0.999
```

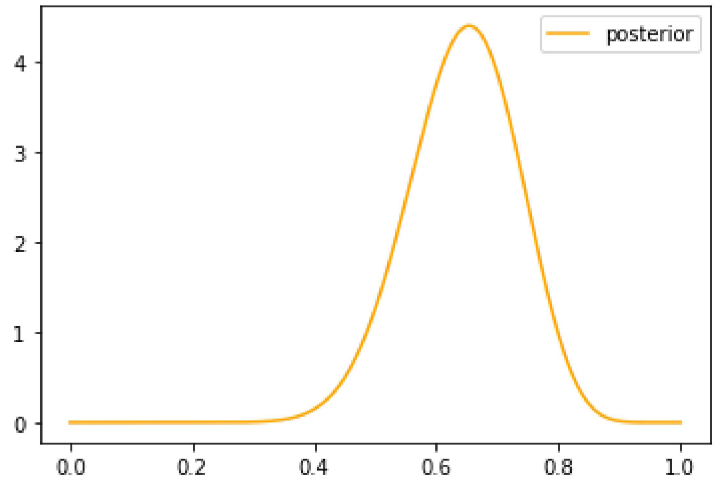
Task 3: Compute $\log p(\theta \mid \mathcal{D}, a, b)$ for different values of θ

The function `loggamma` from the `scipy.special` package might be useful here.

In [12]:
```
def compute_log_posterior(theta, samples, a, b):
    """Compute log p(theta | D, a, b) for the given values of theta.

    Parameters
    ----------
    theta : array, shape (num_points)
        Values of theta for which it's necessary to evaluate the log-prior.
    samples : array, shape (num_samples)
        Outcomes of simulated coin flips. Tails is 1 and heads is 0.
    a, b: float
        Parameters of the prior Beta distribution.

    Returns
    -------
    log_posterior : array, shape (num_points)
        Values of log-posterior for each value in theta.
    """
    ### YOUR CODE HERE ###
    # It is also a Beta function which a = a + T, b = b + H
    num_tail, num_head = helper(samples)
    normalization_constant = loggamma(a + b + num_tail + num_head) - loggamma(
a + num_tail) - loggamma(b + num_head)
    log_posterior = normalization_constant + (a + num_tail -1) * np.log(theta)
+ (b + num_head - 1) * np.log(1 - theta)
    return log_posterior
```

In [13]:
```
x = np.linspace(1e-5, 1-1e-5, 1000)

log_posterior = compute_log_posterior(x, samples, a, b)
posterior = np.exp(log_posterior)
plt.plot(x, posterior, label='posterior', c='orange')
plt.legend()
```

Out[13]: <matplotlib.legend.Legend at 0x1e375a1c908>



Like the prior, the posterior defines a probability distribution over θ and integrates to 1.

In [14]:
```
int_posterior = 1.0 * np.mean(posterior)
print(f'Integral = {int_posterior:.4}')
```

```
Integral = 0.999
```

Task 4: Compute θ_{MLE}

In [15]:
```
num_tail, num_head = helper(samples)
```

In [16]:
```
def compute_theta_mle(samples):
    """Compute theta_MLE for the given data.

    Parameters
    ----------
    samples : array, shape (num_samples)
        Outcomes of simulated coin flips. Tails is 1 and heads is 0.

    Returns
    -------
    theta_mle : float
        Maximum likelihood estimate of theta.
    """
    ### YOUR CODE HERE ###
    return num_tail / (num_head + num_tail)
```

In [17]:
```
theta_mle = compute_theta_mle(samples)
print(f'theta_mle = {theta_mle:.3f}')
```

```
theta_mle = 0.750
```

Task 5: Compute θ_{MAP}

In [18]:
```
def compute_theta_map(samples, a, b):
    """Compute theta_MAP for the given data.

    Parameters
    ----------
    samples : array, shape (num_samples)
        Outcomes of simulated coin flips. Tails is 1 and heads is 0.
    a, b: float
        Parameters of the prior Beta distribution.

    Returns
    -------
    theta_mle : float
        Maximum a posteriori estimate of theta.
    """
    ### YOUR CODE HERE ###
    return (num_tail + a - 1) / (num_tail + a + num_head + b - 2)
```

In [19]:
```
theta_map = compute_theta_map(samples, a, b)
print(f'theta_map = {theta_map:.3f}')
```

```
theta_map = 0.654
```

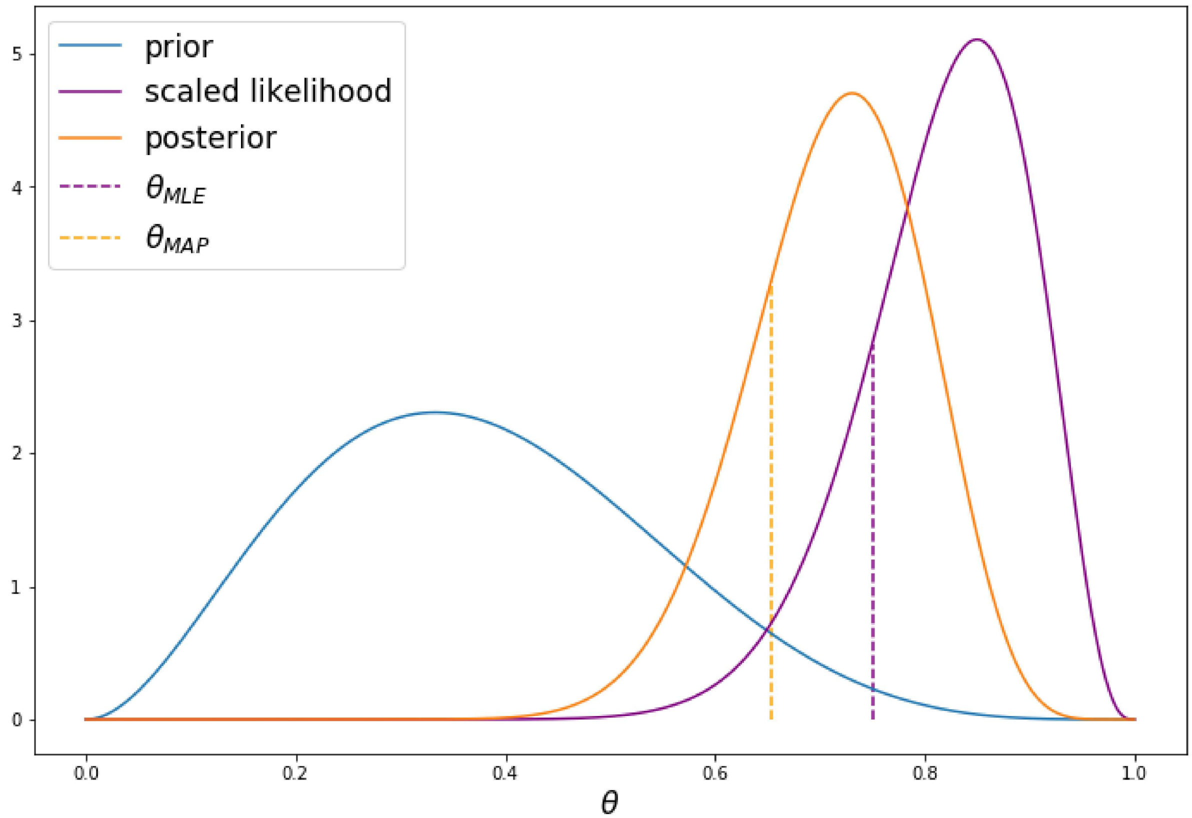
Putting everything together

Now you can play around with the values of `a` , `b` , `num_samples` and `tails_proba` to see how the results are changing.

In [20]:
```
num_samples = 20
tails_proba = 0.7
samples = simulate_data(num_samples, tails_proba)
a, b = 3, 5
print(samples)
```

```
[1 1 1 1 1 1 1 0 0 1 0 1 1 1 1 1 1 1 1 1]
```

In [21]:
```
plt.figure(figsize=[12, 8])
x = np.linspace(1e-5, 1-1e-5, 1000)

# Plot the prior distribution
log_prior = compute_log_prior(x, a, b)
prior = np.exp(log_prior)
plt.plot(x, prior, label='prior')

# Plot the likelihood
log_likelihood = compute_log_likelihood(x, samples)
likelihood = np.exp(log_likelihood)
int_likelihood = np.mean(likelihood)
# We rescale the likelihood - otherwise it would be impossible to see in the p
lot
rescaled_likelihood = likelihood / int_likelihood
plt.plot(x, rescaled_likelihood, label='scaled likelihood', color='purple')

# Plot the posterior distribution
log_posterior = compute_log_posterior(x, samples, a, b)
posterior = np.exp(log_posterior)
plt.plot(x, posterior, label='posterior')

# Visualize theta_mle
theta_mle = compute_theta_mle(samples)
ymax = np.exp(compute_log_likelihood(np.array([theta_mle]), samples)) / int_li
kelihood
plt.vlines(x=theta_mle, ymin=0.00, ymax=ymax, linestyle='dashed', color='purpl
e', label=r'$\theta_{MLE}$')


# Visualize theta_map
theta_map = compute_theta_map(samples, a, b)
ymax = np.exp(compute_log_posterior(np.array([theta_map]), samples, a, b))
plt.vlines(x=theta_map, ymin=0.00, ymax=ymax, linestyle='dashed', color='orang
e', label=r'$\theta_{MAP}$')

plt.xlabel(r'$\theta$', fontsize='xx-large')
plt.legend(fontsize='xx-large')
plt.show()
```


prior
scaled likelihood
posterior
θ_{MLE}
θ_{MAP}
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
θ

In [ ]: