10-424/624: Recitation 3

February 7, 2025

1 The Laplace Approximation

The Beta distribution is given by

$$p(\theta \mid \alpha, \beta) = \frac{1}{\mathcal{B}(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

In this exercise, we derive the Laplace approximation to $\mathcal{B}(\alpha, \beta)$ and analyze its empirical quality. We work with the log of the numerator, represented by the following function:
$\Psi(\theta) = \log\left(\theta^{\alpha-1}(1-\theta)^{\beta-1}\right)$

1. Compute the maximum of $\Psi(\theta)$. What is the value $\hat{\theta}$ for which this maximum is attained?

We have

$$\begin{aligned} \Psi(\theta) &= (\alpha-1)\log\theta + (\beta-1)\log(1-\theta) \\ \implies \Psi'(\theta) &= \frac{\alpha-1}{\theta} - \frac{\beta-1}{1-\theta} \end{aligned}$$

Since $\hat{\theta}$ is the value of θ where $\Psi(\theta)$ is maximum, we must have $\Psi'(\hat{\theta}) = 0$. Thus,

$$\begin{aligned} \frac{\alpha-1}{\hat{\theta}} &= \frac{\beta-1}{1-\hat{\theta}} \\ \hat{\theta} &= \frac{\alpha-1}{\alpha+\beta-2} \end{aligned}$$

The maximum value attained is

$$\begin{aligned} \Psi(\theta)|_{\theta=\hat{\theta}} &= (\alpha-1)\log\left(\frac{\alpha-1}{\alpha+\beta-2}\right) + (\beta-1)\left(\frac{\beta-1}{\alpha+\beta-2}\right) \\ &= (\alpha-1)\log(\alpha-1) + (\beta-1)\log(\beta-1) - (\alpha+\beta-2)\log(\alpha+\beta-2) \\ &= \log\left(\frac{(\alpha-1)^{\alpha-1}(\beta-1)^{\beta-1}}{(\alpha+\beta-2)^{\alpha+\beta-2}}\right) \end{aligned}$$

2. What is the negative of the Hessian of $\Psi(\theta)$ at $\hat{\theta}$?

We have

$$\Psi''(\theta) = -\frac{\alpha-1}{\theta^2} - \frac{\beta-1}{(1-\theta)^2}$$

Plugging in $\hat{\theta}$ for θ, we have

$$\begin{aligned}\Psi''(\theta)|_{\theta=\hat{\theta}} &= -\frac{\alpha-1}{\left(\frac{\alpha-1}{\alpha+\beta-2}^2\right)} - \frac{\beta-1}{\left(\frac{\beta-1}{\alpha+\beta-2}^2\right)} \\ &= -(\alpha+\beta-2)^2\left(\frac{1}{\alpha-1}+\frac{1}{\beta-1}\right) \\ &= -\frac{(\alpha+\beta-2)^3}{(\alpha-1)(\beta-2)}\end{aligned}$$

Thus, we have

$$-\Psi''(\theta)|_{\theta=\hat{\theta}} = \frac{(\alpha+\beta-2)^3}{(\alpha-1)(\beta-1)}$$

3. Using the above two parts, what is the Laplace approximation for $\mathcal{B}(\alpha,\beta)$?

The second-order Taylor series approximation of $\Psi(\theta)$ about $\hat{\theta}$ gives us

$$\begin{aligned}\Psi(\theta) &\approx \Psi(\hat{\theta}) + \frac{1}{2}(\theta-\hat{\theta})^2\Psi''(\hat{\theta}) \\ &= \Psi(\hat{\theta}) - \frac{1}{2}(\theta-\hat{\theta})^2(-\Psi''(\hat{\theta})) \\ &= \log\left(\frac{(\alpha-1)^{\alpha-1}(\beta-1)^{\beta-1}}{(\alpha+\beta-2)^{\alpha+\beta-2}}\right) - \frac{1}{2}(\theta-\hat{\theta})^2\frac{(\alpha+\beta-2)^3}{(\alpha-1)(\beta-1)}\end{aligned}$$

(*Note*: For the sake of compact notation, we use $\Psi''(\theta)|_{\theta=\hat{\theta}} = \Psi''(\hat{\theta})$)
We thus have the following approximation

$$\theta^{\alpha-1}(1-\theta)^{\beta-1} \approx \left(\frac{(\alpha-1)^{\alpha-1}(\beta-1)^{\beta-1}}{(\alpha+\beta-2)^{\alpha+\beta-2}}\right)\exp\left(-\frac{1}{2}(\theta-\hat{\theta})^2\frac{(\alpha+\beta-2)^3}{(\alpha-1)(\beta-1)}\right)$$

Taking the integral w.r.t. θ, our approximation is

$$\begin{aligned}&\int_{-\infty}^{\infty}\left(\frac{(\alpha-1)^{\alpha-1}(\beta-1)^{\beta-1}}{(\alpha+\beta-2)^{\alpha+\beta-2}}\right)\exp\left(-\frac{1}{2}(\theta-\hat{\theta})^2\frac{(\alpha+\beta-2)^3}{(\alpha-1)(\beta-1)}\right)\,d\theta \\ &= \left(\frac{(\alpha-1)^{\alpha-1}(\beta-1)^{\beta-1}}{(\alpha+\beta-2)^{\alpha+\beta-2}}\right)\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}(\theta-\hat{\theta})^2\frac{(\alpha+\beta-2)^3}{(\alpha-1)(\beta-1)}\right)\,d\theta \\ &= \left(\frac{(\alpha-1)^{\alpha-1}(\beta-1)^{\beta-1}}{(\alpha+\beta-2)^{\alpha+\beta-2}}\right)\sqrt{\frac{2\pi(\alpha-1)(\beta-1)}{(\alpha+\beta-2)^3}} \\ &= \sqrt{2\pi}\frac{(\alpha-1)^{\alpha-1/2}(\beta-1)^{\beta-1/2}}{(\alpha+\beta-2)^{\alpha+\beta-1/2}}\end{aligned}$$

4. We now evaluate the quality of our approximation for all combinations of $\alpha \in \{2, 5, 10\}$ and $\beta \in \{2, 5, 10\}$. We calculate the true value of $\mathcal{B}(\alpha, \beta)$, the Laplace approximation, and the relative error.

α	β	True integral	Laplace Approximation	Relative error
2	2	0.166667	0.221557	0.329340
2	5	0.033333	0.036733	0.101986
2	10	0.009091	0.009213	0.013413
5	2	0.033333	0.036733	0.101986
5	5	0.001587	0.001731	0.090475
5	10	0.000100	0.000105	0.051766
10	2	0.009091	0.009213	0.013413
10	5	0.000100	0.000105	0.051766
10	10	0.000001	0.000001	0.041004

We observe that the relative error decreases with increasing α and β.

2 A Valid Kernel

1. Suppose you have a finite input space $\mathcal{X}$ and consider the following simple kernel function over this space:

$$k(x, x') = \begin{cases} 1 \text{ if } x = x' \\ 0 \text{ otherwise.} \end{cases}$$

Prove this is a legal kernel by describing an implicit mapping $\phi : \mathcal{X} \to \mathbb{R}^D$ (for some value D) such that $k(x, x') = \phi(x)^T \phi(x')$.

> Let ϕ be a mapping such that $\phi(x) \in \mathbb{R}^{|X|}$. Then there is a component for every element $x' \in X$. We set that component of $\phi(x)$ equal to 1 if $x = x'$ and otherwise we set that component equal to 0. Therefore $\phi(x)^T \phi(x') = 1$ if $x = x'$ and otherwise it equals 0. This means that $k(x, x') = \phi(x)^T \phi(x')$.

2. Describe the behavior of this kernel and why it might not make for a good choice in the context of machine learning. Frame your answer in terms of the generalization capabilities of a model that uses the implicit feature transformation you identified, ϕ.

> This kernel is effectively memorizing each of the (finitely-many) possible inputs and mapping them to an index or dimension in the transformed feature space. Models trained using this kernel are highly liable to overfit given the complexity of the implied transformation.

3 Bayesian Model Selection

(from Information Theory, Inference, and Learning Algorithms by David MacKay)
Random variables x come independently from a probability distribution $P(x)$. According to model $\mathcal{H}_0$, $P(x)$ is a uniform distribution

$$P(x|\mathcal{H}_0) = \frac{1}{2}, \quad x \in (-1, 1)$$

According to model $\mathcal{H}_1$, $P(x)$ is a nonuniform distribution with an unknown parameter $m \in (-1, 1)$:

$$P(x|m, \mathcal{H}_1) = \frac{1}{2}(1 + mx), \quad x \in (-1, 1)$$

Given the data $D = \{0.3, 0.9\}$, what is the evidence for $\mathcal{H}_0$ and $\mathcal{H}_1$? In other words, what is $P(D|\mathcal{H}_0)$ and $P(D|\mathcal{H}_1)$?

$$p(D = \{0.3, 0.9\}|\mathcal{H}_0) = \left(\frac{1}{2}\right)^2 = \frac{1}{4}$$

We want $P(D|\mathcal{H}_1)$ and have $P(D|m, \mathcal{H}_1)$ and $P(m|\mathcal{H}_1)$
Note that $\int P(D|m, \mathcal{H}_1)P(m|\mathcal{H}_1)\ dm = P(D|\mathcal{H}_1)$

Plugging in:

$$\begin{aligned}
&\int_{-1}^{1} (\frac{1}{2}(1 + 0.3m))(\frac{1}{2}(1 + 0.9m)) * \frac{1}{2}\ dm \\
&= \frac{1}{8}\int_{-1}^{1} (1 + 0.3m)(1 + 0.9m)\ dm \\
&= \frac{1}{8}\int_{-1}^{1} (1 + 0.27m^2 + 1.2m)\ dm \\
&= \frac{1}{8}\left[m + 0.09m^3 + 0.6m^2\right]_{-1}^{1} \\
&= \frac{1}{8}\left[1.69 - (-0.49)\right] \\
&= \frac{2.18}{8}
\end{aligned}$$

4 Plotting Tutorial

Throughout this course, you will be exposed to and expected to generate beautiful figures such as those below. We will be heavily utilizing the Python package 'matplotlib'.

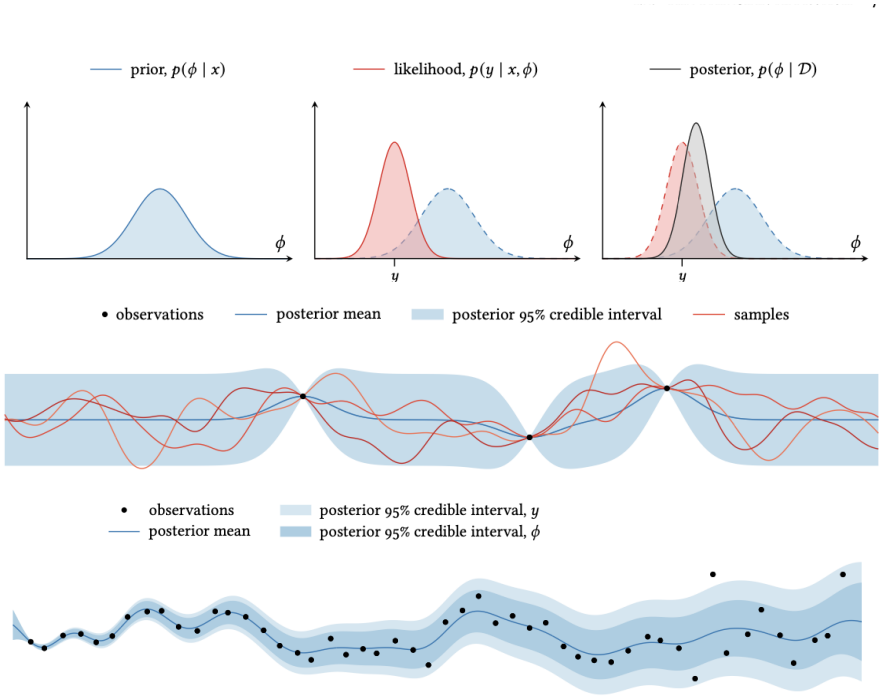


Figure 1: Collection of visualizations from the field of Bayesian machine learning. Credit: Roman Garnett

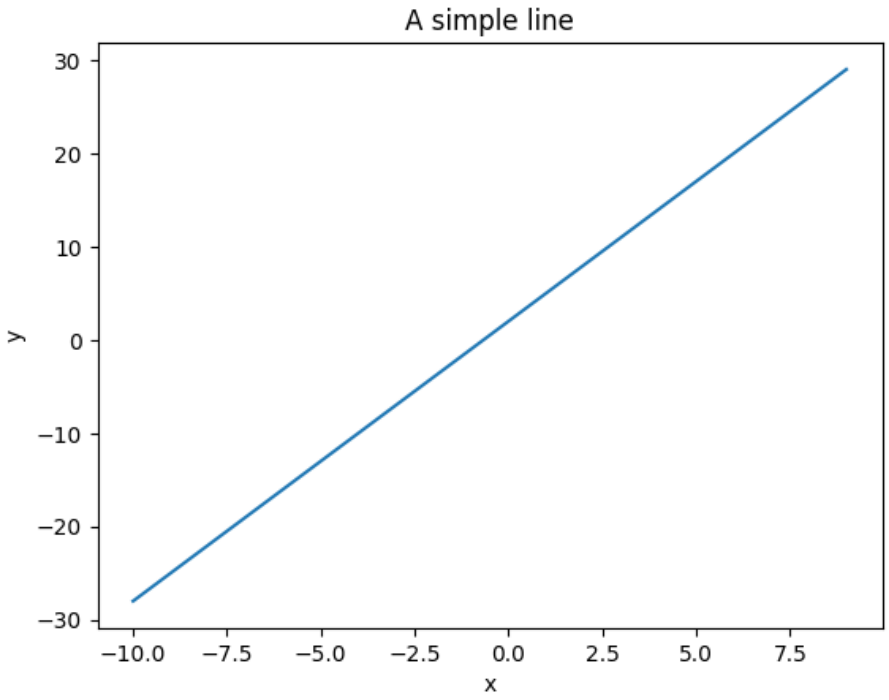
Matplotlib does not allow directly visualizing functions – instead you must plot a series of points. In order to visualize a function, you can generate a regular interval of points along the x-axis using `np.arange`, `np.linspace`, or similar and plot the function output on the y-axis.

Please complete the following pseudocode to plot a line with equation $y = 3x + 2$ in the range [-10, 10].

```
import numpy as np
import matplotlib.pyplot as plt

x =                          #TODO: Fill this in
y =                          #TODO: Fill this in

plt.plot(x, y)
plt.xlabel("x")
plt.ylabel("y")
plt.title(f"A simple line")
plt.show()
```

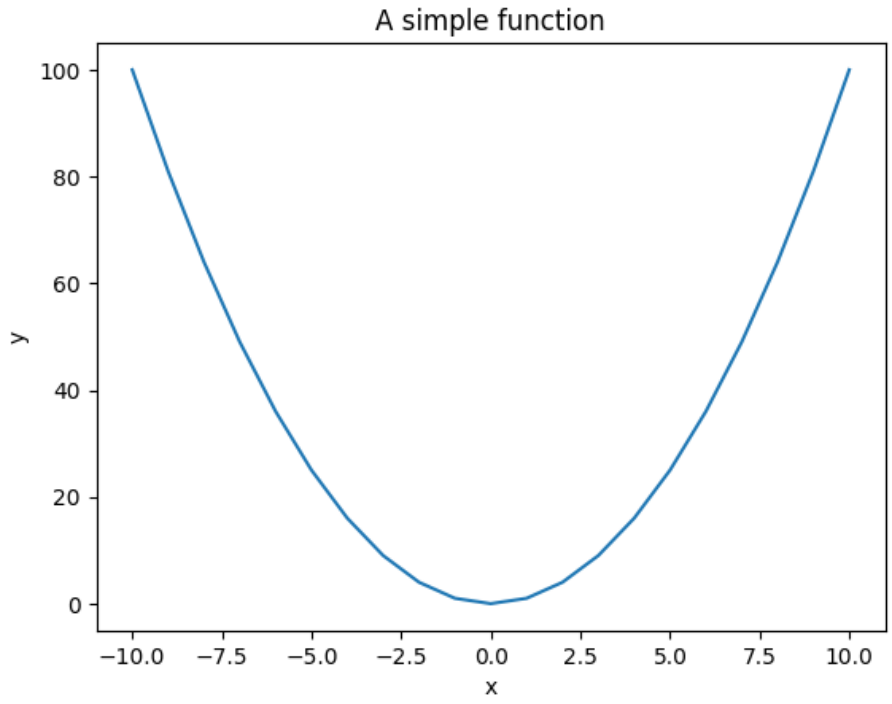


Next, visualize the function $y = x^2$ in the range [-10, 10]

```
import numpy as np
import matplotlib.pyplot as plt

x =                          #TODO: Fill this in
y =                          #TODO: Fill this in

plt.plot(x, y)
plt.xlabel("x")
plt.ylabel("y")
plt.title(f"A simple function")
plt.show()
```



Finally, you can use the command `plt.fill_between` in order to visualize a credible interval. Please read the documentation to complete the starter code below:

`https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.fill_between.html`

```
import numpy as np
import matplotlib.pyplot as plt

x =                          #TODO: Fill this in
y =                          #TODO: Fill this in
std_dev = 10

plt.plot(x, y)
plt.fill_between(
                              #TODO: Fill this in
                              #TODO: Fill this in
                              #TODO: Fill this in
    color="lightblue",
    label="Example Confidence Interval",
)

plt.xlabel("x")
plt.ylabel("y")
plt.title(f"A simple function")
plt.legend()
plt.show()
```