Bayesian Reasoning

Consider the following model:
\[ y_i = \hat{\beta}_0 + x_i \hat{\beta}_1 + x_i^2 \hat{\beta}_2 + \epsilon_i \]
where the probabilistic view of linear regression:
\[ y_i = \beta_0 + x_i \beta_1 + x_i^2 \beta_2 + \epsilon_i \sim N(0, \sigma^2) \]

We are able to estimate \( \beta_0, \beta_1, \beta_2 \) using MLE estimation. We assume that every point has the same variance. And we are able to compute that variance using MLE.

The ability to model uncertainty is very important. It also enables us to do the exploration vs. exploitation trade-off.

Example: drilling for minerals.

Bayesian learning allows us to do just that.
Problem: A doctor has good news and a bad news for a patient.

Bad News: The patient has just tested positive for a serious disease. The test is 99% accurate.

Good News: This is a very rare disease. Only 1 in 10,000 people have it.

Question: Should the patient be very worried or not?

* This kind of question arise in many different areas: legal proceedings, sports doping, etc.

Bayes' Rule
- A rule about how humans reason
- Ted Talk by Alex Gopnik (Psychologist at Berkeley)
- This rule allows us to update probabilities.

\[ P(A|B) = \frac{P(B|A)P(A)}{P(B)} \]

- Very useful for the universal problems: computer vision
- Key to perception.

\[ P(AB) = P(B|A)P(A) = P(A|B)P(B) \]

- Joint \( \rightarrow \) marginal

\[ P(B|A) : \text{prob. of } B \text{ given } A. \]
\[ \int P(AB) \, dA \, dB = 1 \]
\[ \int P(A|B) \, dA = 1 \]

This is a distribution over \( A \) when \( B \) is given.

\[ \int P(A) \, dA = 1 \]
\[ \int P(AB) \, dB = P(A) \quad \text{integrating with} \quad B. \]

This is also referred to as marginalization in the world of probability.

Learning and Bayesian Inference:

\[ p(h|d) = \frac{p(d|h) \, p(h)}{\sum_{h'} p(d|h') \, p(h')} \]

\( p(h) \) refers to prior belief.
\( p(d|h) \) refers to likelihood.
\( p(h|d) \) refers to posterior.

This integral is very hard to do in practice. We can do it for \( 3 \) Gaussians though.
Let's use Bayes' rule to see if the patient should be worried:

1. Test is 99.7% accurate: \( P(T=1 \mid D=1) = 0.99 \) and \( P(T=0 \mid D=0) = 0.99 \)

2. Disease affects 1 in 10,000: \( P(D=1) = 0.0001 \)

\[
P(D=1 \mid T=1) = \frac{P(T=1 \mid D=1) P(D=1)}{P(T=1 \mid D=0) P(D=0) + P(T=1 \mid D=1) P(D=1)}
\]

\[
= \frac{(0.99)(0.0001)}{(1-0.99)(1-0.0001) + (0.99)(0.0001)}
\]

\[
= 0.9898
\]

This is the probability of the patient having the disease given that the test came positive.

Example:

\[
P(\text{words} \mid \text{sounds}) \propto P(\text{sound} \mid \text{words}) P(\text{words})
\]

Bayesian Learning for Model Parameters

1. Given in data \( D = \{ x^{(1)}, x^{(2)}, \ldots, x^{(n)} \} \), write down the likelihood of data \( P(D \mid \theta) \)
2. Specify prior \( P(\theta) \)
3. Compute posterior \( P(\theta \mid D) = \frac{P(D \mid \theta) P(\theta)}{P(D)} \)
Prior \( p(\theta) \) encodes our belief about the parameters.

We are assuming that \( \theta \) is a random variable. Note that for MLE, \( \theta \) is not a random variable.

The prior can also be viewed as my initial belief.

\[
p(D|\theta) = \frac{p(\theta) p(D)}{p(D)} \propto p(D|\theta) p(\theta)
\]

This can be seen as simply normalizing \( p(D|\theta)p(\theta) \).

* Frequentist vs. Bayesian: Bayesian claim that it is possible to assign probabilities with even seeing the frequencies of events in question.

**Bayesian Linear Regression**

1. The likelihood is Gaussian \( N(y | x\theta, \sigma^2 I_n) \)

The conjugate prior is also a Gaussian \( p(\theta) = N(\theta | \theta_0, V_0) \). \( \theta_0 \) is the mean and \( V_0 \) is the variance.

\[
p(\theta | x, y, \sigma^2) \propto N(\theta | \theta_0, V_0) N(y | x\theta, \sigma^2 I_n) = N(\theta | \theta_n, V_n)
\]

\[
\theta_n = V_n V_0^{-1} \theta_0 - \frac{1}{\sigma^2} V_n x^T y
\]

\[
V_n^{-1} = V_0^{-1} + \frac{1}{\sigma^2} x^T x
\]

This sort of calculations are also called "completing the squares" or conjugate analysis. Both prior and the posterior has the same shape.

* You need a course on "Bayesian Analysis!"
\[ p(\theta | y, x, \sigma^2) \propto p(y | x, \theta, \sigma^2) p(\theta) \]

\[ \propto e^{-\frac{1}{2\sigma^2} \left( y - \theta x \right)^T (y - \theta x)} e^{\frac{1}{2} (\theta - \theta_0)^T V_0^{-1} (\theta - \theta_0)} \]

\[ = e^{\frac{1}{2} \sum \left( y_i - \theta_0 \theta_0^T \right)^T (y_i - \theta_0 \theta_0^T) + \left( \theta - \theta_0 \right)^T V_0^{-1} (\theta - \theta_0) \frac{1}{2}} \]

\[ = e^{-\frac{1}{2} \left( \theta - \theta_0 \right)^T \left( \sigma^2 \theta_0 \theta_0^T + \sigma^2 V_0 \right) \theta + \left( \theta - \theta_0 \right)^T V_0^{-1} (\theta - \theta_0) \frac{1}{2}} \]

\[ = e^{\frac{1}{2} \sum \left( \sigma^2 \theta_0 \theta_0^T + \sigma^2 V_0 \right) \theta + \left( \theta - \theta_0 \right)^T V_0^{-1} (\theta - \theta_0) \frac{1}{2}} \]

\[ = e^{\frac{1}{2} \sum \left( \sigma^2 \theta_0 \theta_0^T + \sigma^2 V_0 \right) \theta + \left( \theta - \theta_0 \right)^T V_0^{-1} (\theta - \theta_0) \frac{1}{2}} \]

\[ = e^{-\frac{1}{2} \theta^T \left( \sigma^2 \theta_0 \theta_0^T + \sigma^2 V_0 \right) \theta + \left( \theta - \theta_0 \right)^T V_0^{-1} (\theta - \theta_0) \frac{1}{2}} \]

\[ = e^{\frac{1}{2} \theta^T \left( \sigma^2 \theta_0 \theta_0^T + \sigma^2 V_0 \right) \theta + \left( \theta - \theta_0 \right)^T V_0^{-1} (\theta - \theta_0) \frac{1}{2}} \]

\[ = e^{-\frac{1}{2} \theta^T \left( \sigma^2 \theta_0 \theta_0^T + \sigma^2 V_0 \right) \theta + \left( \theta - \theta_0 \right)^T V_0^{-1} (\theta - \theta_0) \frac{1}{2}} \]

\[ = e^{\frac{1}{2} \theta^T \left( \sigma^2 \theta_0 \theta_0^T + \sigma^2 V_0 \right) \theta + \left( \theta - \theta_0 \right)^T V_0^{-1} (\theta - \theta_0) \frac{1}{2}} \]

\[ = e^{\frac{1}{2} \theta^T \left( \sigma^2 \theta_0 \theta_0^T + \sigma^2 V_0 \right) \theta + \left( \theta - \theta_0 \right)^T V_0^{-1} (\theta - \theta_0) \frac{1}{2}} \]
Bayesian Linear Regression.

The likelihood is Gaussian: \( N(y|X\theta, \sigma^2 I_n) \)

The conjugate prior is also Gaussian: \( p(\theta) = N(\theta|\theta_0, V_0) \)

mean: \( \theta_0 \)

variance: \( V_0 \)

Later you'll notice that if we make \( \theta_0 = 0 \) and \( V_0 \)

equal to a diagonal matrix, what you get is

a ridge regression, i.e., Bayesian linear

regression will subsume ridge regression.

For standard Bayesian linear regression, we will not compute a single value of \( \theta \). Instead we will estimate a distribution over \( \theta \). Specifically we want to compute the posterior \( p(\theta|x,y,\sigma^2) \)

\[ p(\theta|x,y,\sigma^2) \propto N(\theta|\theta_0, V_0) \cdot N(y|X\theta, \sigma^2 I) = N(\theta|\theta_n, V_n) \]

\( \theta_n \): mean

\( V_n \): variance, that models the uncertainty.

Through conjugate analysis or completing squares exercise, we can find out the value \( \theta_n \) for \( \theta_n \) and \( V_n \).

\[ \theta_n = V_n V_0^{-1} \theta_0 - \frac{1}{\sigma^2} V_n X^T y \]

\[ V_n^{-1} = V_0^{-1} - \frac{1}{\sigma^2} X^T X \]

Conjugate analysis: what conjugate analysis means that the prior and the posterior has the shape.

* You need a course on Bayesian statistics."
We are exploiting conjugate analysis. That is to say we are picking a prior such that our posterior has the same shape as the prior. Ideally I would like the freedom to pick any prior; however, that makes analysis/computation for posterior very difficult.

Conjugate analysis: assume \( \sigma^2 \) is known.

\[
P(\theta | y, x, \sigma^2) \propto P(y | x, \sigma^2) P(\theta|\sigma^2)
\]

\[
\propto e^{-\frac{1}{2}(\theta - \theta_0)^T V_\sigma^{-1} (\theta - \theta_0)} e^{-\frac{1}{2} \theta^T (B - B_0) V_0^{-1} (B - B_0) \theta}
\]

Let's combine these two terms and complete squares.

Proportionality allows us to get rid of constant:

\[
\begin{align*}
\mathcal{L} &= e^{-\frac{1}{2} \theta^T (B - B_0) V_0^{-1} (B - B_0) \theta} e^{-\frac{1}{2} (Y - XB)^T (Y - XB) - \frac{1}{2} \theta^T (B - B_0) V_0^{-1} (B - B_0) \theta} \\
&= e^{-\frac{1}{2} \theta^T (B - B_0) V_0^{-1} (B - B_0) \theta} e^{\theta^T V_0^{-1} \theta + \theta^T V_0^{-1} \theta} \\
&= e^{\theta^T V_0^{-1} \theta}
\end{align*}
\]

\[
\begin{align*}
\mathcal{L} &= \frac{1}{\mathcal{L}_0} (\theta^T (B^2) \theta + \theta^T (\theta^2 I)^T \theta + V_0^{-1} \theta) \\
&= \frac{1}{\mathcal{L}_0} \theta^T V_0^{-1} \theta + \theta^T (\theta^2 I)^T \theta + V_0^{-1} \theta
\end{align*}
\]

\[
\begin{align*}
\mathcal{L} &= \frac{1}{\mathcal{L}_0} \theta^T V_0^{-1} \theta + \theta^T (\theta^2 I)^T \theta + V_0^{-1} \theta \\
&= \frac{1}{\mathcal{L}_0} \theta^T V_0^{-1} \theta + \theta^T (\theta^2 I)^T \theta + V_0^{-1} \theta
\end{align*}
\]

\[
\mathcal{L} = \frac{1}{\mathcal{L}_0} \theta^T V_0^{-1} \theta + \theta^T (\theta^2 I)^T \theta + V_0^{-1} \theta
\]

We want to get rid of this term.

If we choose \( B_0 = 0 \) and \( V_0 = \sigma^2 I_d \), which is a spherical Gaussian prior, the posterior reduces to
\[ \hat{\theta}_n = \frac{1}{\sigma^2} V_n X^T y = \frac{1}{\sigma^2} \left( \frac{1}{\tau_0^2} I + \frac{1}{\sigma^2} X^T X \right)^{-1} X^T y \]

where \( \chi = \sigma^2 / \tau_0^2 \). We just recovered ridge regression.

Also if you make your prior flat, you get maximum likelihood estimate.

Aside,

let set \( \hat{\theta}_n X V_n^{-1} \frac{X^T X}{\sigma^2} - \hat{\theta}_0 V_n^{-1} = 0 \) and solve for \( \hat{\theta}_n \). This yields

\[ \hat{\theta}_n = V_n \left[ \frac{V_n^{-1} \hat{\theta}_0 + \frac{X^T y}{\sigma^2}}{\sigma^2} \right] \]

And when this happens, we get

\[ P(\theta | x, y, \sigma^2) \propto e^{-\frac{1}{2} (\theta - \hat{\theta}_n)^T V_n^{-1} (\theta - \hat{\theta}_n)} \]

By the definition of multivariate Gaussian, we have

\[ \int e^{-\frac{1}{2} (\theta - \hat{\theta}_n)^T V_n^{-1} (\theta - \hat{\theta}_n)} \, d\theta = \sqrt{2\pi \det V_n} \]

\[ \therefore P(\theta | x, y, \sigma^2) = \frac{1}{\sqrt{2\pi \det V_n}} e^{-\frac{1}{2} (\theta - \hat{\theta}_n)^T V_n^{-1} (\theta - \hat{\theta}_n)} \]

You can easily derive this integral from first principles.

Q. So what happens if we have a prior that is not amenable to conjugate analysis? Say I do not know the prior and I picked a uniform distribution.

A. If we don't know the shape of the posterior, we will have to use numerical techniques for evaluating the integral to find the posterior. We will use for example Monte Carlo techniques.
In this derivation of prior we assumed $\sigma^2$ is known. We can also discuss a case where a prior on variance is given. There for conjugate analysis we’ll use the inverse Wishart distribution.

* ML: A probabilistic perspective. Ch. 5

A Theorem for Gaussians *(Kevin Murphy’s Book)*

$$p(x) = \mathcal{N}(x|\mu_x, \Sigma_x) \quad \text{marginal}$$

$$p(y|x) = \mathcal{N}(y|Ax + b, \Sigma_y) \quad \text{likelihood}$$

$$p(x|y) = \mathcal{N}(x|\mu_{x|y}, \Sigma_{x|y})$$

$$\Sigma_{x|y}^{-1} = \Sigma_x^{-1} + A^T \Sigma_y^{-1} A$$

$$\mu_{x|y} = \Sigma_{x|y} \left[ A^T \Sigma_y^{-1} (y - b) + \Sigma_x^{-1} \mu_x \right]$$

$$p(y) = \mathcal{N}(y|A\mu_x + b, \Sigma_y + A \Sigma_x A^T) \quad \text{3}$$

The above theorem holds for any two variables $x$ and $y$.

When we were doing the completing squares exercise, I was actually trying to prove box 3.

Recall that we need 1, 2, and 3 above to apply the Bayes Rule,

$$p(x|y) = \frac{p(y|x) \ p(x)}{p(y)}$$

Aside: The 3 box is often called “convolution”. That’s because $p(y) = \int p(y|x') \ p(x') \ dx'$.

* chapter 4 of Kevin’s book.
Bayesian vs. ML plugin prediction

Posterior mean: \( \hat{\theta}_n = (\lambda I_d + X^T X)^{-1} X^T y \)

Posterior variance: \( \nu_n = \sigma^2 (\lambda I_d + X^T X)^{-1} \)

To predict, Bayesians marginalize over the posterior.

Let \( x^* \) be a new input. Then the prediction, given the training data \( D = (X, y) \), is:

\[
P(y|x^*, \sigma^2) = \int N(y|x^T \theta, \sigma^2) N(\theta|\theta_n, \nu_n) \, d\theta
\]

\[
= N(y|x^T \theta_n, \sigma^2 + x^T \nu_n x^*)
\]

Or, for each possible value of \( \theta \), the prediction is weighted by the posterior. So it is a weighted prediction. It is weighted over an infinite domain.

The frequentists tend to model the prediction using the likelihood.

* Each \( \theta \) gets weighted by its posterior probability.

This is an example of an ensemble predictor. In contrast an ML predictor is:

\[
P(y|x^*, \sigma^2) = N(y|x^T \theta_M, \sigma^2)
\]

This assumes that there is one \( \theta \).

ML simply computes (re-writes) (1) above as follows:

\[
N(y|x^T \theta, \sigma^2) N(\theta|\theta_M, \nu_M) \, d\theta
\]

\[
= N(y|x^T \theta, \sigma^2) \delta(\theta) \, d\theta
\]

\( \delta(\theta) \) is just a spike at one place.

Also called Dirac fn. or impulse fn.

Integral w.r.t. Delta picks \( \theta_M \).
* ML assumes there's in only one \( \theta \).

Bayesians assume there are infinite \( \theta \)s.

Having said above \( \theta_n \) and \( \theta_m \) might be quite similar bearing any numerical \( \theta \) arises. However, the term \( \alpha^T \hat{V}_n \hat{x}^* x^T V_n x^* \) is important, which only appears in Bayesian prediction. This term gives us confidence intervals.