Week 5: Gaussian Posterior
Machine Learning
CS1675
Fall 2023

February 7, 2024

First, let us define our normal (Gaussian) distribution. For a given distribution with mean μ and variance σ^2, the probability of observing a value x drawn from that distribution is given by:

$$\texttt{normal}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{1}$$

In Bayesian inference, we seek to estimate some unknown parameters — μ and σ^2 for the Gaussian — based on some observed data points $\mathbf{x} = [x_1, x_2, \ldots, x_N]$ and a prior assumption about the distribution of the unknown parameters. For this document, we will assume that x_n is a real-valued number drawn from a normal distribution with unknown mean μ and known variance σ^2. Our goal is to find the *maximum a posteriori* (MAP) estimate of μ, μ_{MAP}. This is similar to the maximum likelihood estimate (MLE), except that MAP maximizes the joint probability of the likelihood and prior, while MLE maximizes only the likelihood.

Recall that the posterior is proportional to the likelihood and prior. Our likelihood is a normal distribution, and the conjugate prior of a normal distribution is also normal. Because of this choice of prior, the posterior will also be normal.

$$p(\mu|\mathbf{x}, \sigma^2) \propto \prod_{n=1}^{N} \left\{ \exp\left(-\frac{1}{2\sigma^2}(x_n - \mu)^2\right) \right\} \times \exp\left(-\frac{1}{2\tau_0^2}(\mu - \mu_0)^2\right)$$

Thus our posterior is proportional to a product of exponential terms. Using the identity $e^a e^b = e^{a+b}$, and ignoring the normalizing factor $(1/\sqrt{2\pi\sigma^2})(1/\sqrt{2\pi\tau^2})$ we can rewrite the above equation as:

$$p(\mu|\mathbf{x},\sigma^2) \propto \exp\left(\frac{-\sum_N (x_n-\mu)^2}{2\sigma^2} - \frac{(\mu-\mu_0)^2}{2\tau_0^2}\right) \tag{2}$$

Our goal is now to rearrange the terms in (2) so that we can solve for the posterior mean, μ_{post}. In other words, we will use algebra to transform the previous equation to the form:

$$\texttt{normal}(\mu|\mathbf{x},\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(\mu-\mu_{post})^2}{\sigma^2}\right)$$

We will be dropping constant terms (those which do not include μ) because we can use the above exponential identity and account for it in our normalizing constant. Let us just focus on the exponential power for now from (2)

$$\begin{aligned}
\left(\frac{-\sum_N (x_n-\mu)^2}{2\sigma^2} - \frac{(\mu-\mu_0)^2}{2\tau_0^2}\right) \\
-\frac{1}{2}\left(\frac{\sum_N (x_n-\mu)^2}{\sigma^2} + \frac{(\mu-\mu_0)^2}{\tau_0^2}\right) \\
-\frac{1}{2}\left(\frac{\sum_N x_n^2 - 2\mu\sum_N x_n + \sum_N \mu^2}{\sigma^2} + \frac{\mu^2 - 2\mu\mu_0 + \mu_0^2}{\tau_0^2}\right)
\end{aligned}$$

That last step is just distributing the sum over each term in the expansion of $(x_n - \mu)^2$. The first sum can be ignored because it doesn't depend on μ. The second sum can be rewritten with the identity $\frac{1}{N}\sum_N x_n = \bar{\mathbf{x}}$

$$\begin{aligned}
-\frac{1}{2}\left(\frac{-2\mu N\bar{\mathbf{x}} + N\mu^2}{\sigma^2} + \frac{\mu^2 - 2\mu\mu_0 + \mu_0^2}{\tau_0^2}\right) \\
-\frac{1}{2}\left(\frac{-2\mu N\bar{\mathbf{x}}\tau_0^2 + N\mu^2\tau_0^2 + \mu^2\sigma^2 - 2\mu\mu_0\sigma^2 + \mu_0^2\sigma^2}{\sigma^2\tau_0^2}\right) \\
-\frac{1}{2}\left(\frac{\mu^2(N\tau_0^2+\sigma^2) - 2\mu(N\bar{\mathbf{x}}\tau_0^2 + \mu_0\sigma^2)}{\sigma^2\tau_0^2}\right)
\end{aligned}$$

Next, we will isolate μ^2 in the numerator:

$$-\frac{1}{2}\left(\frac{\mu^2 - 2\mu\frac{N\bar{\mathbf{x}}\tau_0^2+\mu_0\sigma^2}{N\tau_0^2+\sigma^2}}{\frac{\sigma^2\tau_0^2}{N\tau_0^2+\sigma^2}}\right)$$

The next step may be a little unusual. Remember how we dropped terms that did not rely μ because the law of exponents allows us to account for it in the normalizing constant? Now we are going to do the reverse and add our own term, with the same justification of the normalization constant. Imagine that the numerator in the previous equation had a third term which is equal to the square of the fraction after -2μ. Then we could express the entire numerator as a square of differences, leaving us with a final expression for the exponent in our normally-distributed posterior:

$$p(\mu|\mathbf{x}, \sigma^2, \mu_0, \tau_0^2) \propto \exp\left(-\frac{1}{2}\left[\frac{\left(\mu - \frac{N\bar{\mathbf{x}}\tau_0^2 + \mu_0\sigma^2}{N\tau_0^2 + \sigma^2}\right)^2}{\frac{\sigma^2\tau_0^2}{N\tau_0^2 + \sigma^2}}\right]\right)$$

By comparison to (1), we see that the posterior is itself a normal distribution with mean

$$\mu_{post} = \frac{N\bar{\mathbf{x}}\tau_0^2 + \mu_0\sigma^2}{N\tau_0^2 + \sigma^2} = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{N}{\sigma^2}\bar{\mathbf{x}}}{\frac{1}{\tau_0^2} + \frac{N}{\sigma^2}}$$

Likewise, the posterior variance can be expressed as:

$$\sigma_{post}^2 = \frac{\sigma^2\tau_0^2}{N\tau_0^2 + \sigma^2}$$

Thus, we have succeeded in expressing our posterior probability in terms of the likelihood ($\texttt{normal}(\mathbf{x}|\mu, \sigma^2)$) and prior distributions ($\texttt{normal}(\mu|\mu_0, \tau^2)$). We can do more algebra to get these into the same forms as the lecture slides.

This proof is based on the document at this link.