Renormalization Group and Deep Neural Networks

Liam Damewood

Bayesian Statistics

This is a brief intro to Bayesian statistics with examples. The Bayesian method compares the probabilities of models M based on the available data D. Before collecting data, there may be some prior belief about the distribution of data. This is the prior probability distribution $P(D)$. Given some model M, there is an associated probability of getting data D. This is the support for M given D, or $P(D|M)/P(M)$. After collecting data D, the support for model M may increase of decrease and the probability that the model M is supported by data D is $P(M|D)$, the posterior probability distribution.

Bayes' theorem states that

$$P(M|D) = \frac{P(M)P(D|M)}{P(D)}.$$

Biased coin example

A biased coin provides an easy example of using Bayes' theorem. We will assign a hyperparameter b, which describes the amount of bias in the coin, such that the support for the model with parameter b is

$$P(1|b)/P(1) = b$$
$$P(0|b)/P(0) = 1 - b$$

where 1 and 0 represent Heads and Tails, respectively.

A fair coin will have $b = 0.5$ so the probabilities of heads or tails is 50% each. Also note that $P(1) + P(0) = 1$. Without collecting any data (flipping the biased coin multiple times), we do not have a clear indication what b is. We might assume that $b = 0.5$, but instead, let's assume that b can be any value, so the prior belief distribution is

$$P(b) = 1$$

After flipping the coin once, let's assume it comes up Heads (1). Using Bayes' theorem, the posterior probability is

$$P(b|1) = b$$

After N flips, we will have N_h heads and N_t tails, so the probability of the model is

$$P(b|N_h heads, N_t tails) = b^{N_h}(1-b)^{N_t}$$

Polynomial fit example

Fitting data to polynomials is pretty straightforward using linear least squares. Given a set of data $D = (x_0, y_0), ..., (x_N, y_N)$, the fit can easily be obtained by solving for the coefficients v in the Normal equation:

$$min_v ||Av - b||^2$$

thus

$$v = (A'A)^{-1}A'v$$

where A' denotes the transpose.

Using Bayesian statistics, the Normal equation can be derived using the prior belief that any parameters v can fit the data well, so that P(v) = 1. The normal equation solves for the parameters v that maximize the likelihood of $P(v|D)$. Maximizing the likelihood of the posterior distribution is equivalent to minimizing the negative log of the distribution. Taking the negative log of Bayes' equation results in

$$-logP(v|D) = -logP(v) - logP(D|v) + logP(D)$$

and then we want to find where this is minimized so we take the derivative with respect to the parameters v and set it to zero. The prior probability is one, so $\log(1) = 0$ and the last term does not involve v so it drops out. Minimizing the negative log probability in this case means that we need to maximize the probability of getting the data given the parameters v, or the support. If we assume that the data fits the model v but with added Gaussian noise, then

$$P(D|v) = e^{-(Av-b)^2/2\sigma_r}$$

so we are inclined to minimize the term in the exponent to achieve maximum probability. The variance of the residual data $\sigma_y = var(Av - b)$ is a constant, so we arrive back to the Normal equation.

The normal equation assumed that the prior probability allowed equal probability for all models parameterized by v. Instead, if we have some prior belief about the distribution of v, then we can work that into the Normal equation. If our prior belief is expressed as a gaussian characterized by some matrix Γ (the Tikhonov matrix)

$$P(v) = e^{-(\Gamma v)^2/2\sigma_v}$$

where σ_v is the variance in the parameters v, the Normal equation becomes the regularized Normal equation:

$$\min_v ||Av - b||^2/2\sigma_r + ||\Gamma v||^2/2\sigma_v$$

thus

$$v = (A'A + \Gamma'\Gamma\sigma_r/\sigma_v)^{-1}A'b.$$

One particular solution to the regularized Normal equation is when Γ is proportional to the identity matrix $\Gamma = \alpha I$. This choice tries to minimize the residual error $(Av - b)$ but not at the cost of making the parameters v too large.