Machine Learning I

Lecture 20: The B Hive: Bootstraping, Bagging, Bumping and Bayesianism

Nathaniel Bade

Northeastern University Department of Mathematics

Table of contents

1. Bootstraping Confidence Intervals

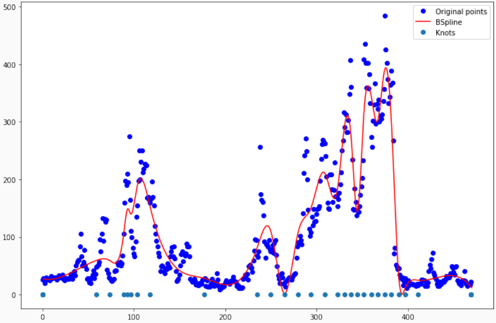
2. Parametric Bootstrapping and Maximum Likelihood

3. Nonparametric Bootstraping and Bayesian Decision Making

4. Bagging and Bumping

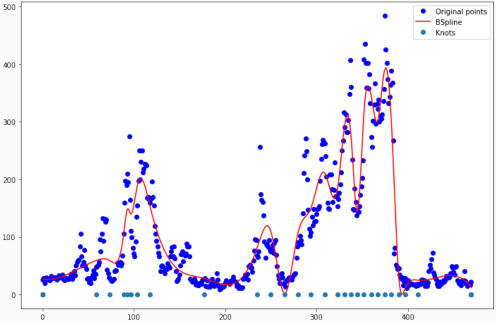
Bootstraping Confidence Intervals

Bootstrapping



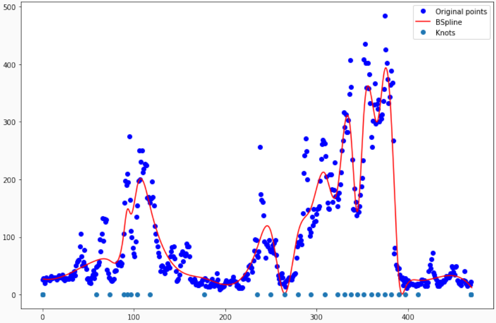
Bootstrapping is the process of resampling a set of training data without replacement and using the newly generated datasets to derive information about our models. For example, suppose we fit some noisy data with a smoothing spline as above.

Bootstrapping



How accurate is the spline fit above? One of the ways to assess accuracy is to construct a confidence interval using the covariance and standard error.

Bootstrapping



Writing the fitting in basis notation,

$$f(x) = \sum_{j=k}^{K} \beta_k h_k(x) \, ,$$

we can think of $f(x)$ as estimating $E(Y|X = x)$.

Bootstrapping

Using our linear methods, we can then write $\mathbf{H}_{ij} = h_j(x_i)$ and as usual

$$\mathrm{RSS} = \sum_{i=1}^{N} \big(y - f(x_i)\big)^2 = (\mathbf{y} - \mathbf{H}\beta)^T (\mathbf{y} - \mathbf{H}\beta)$$

can be minimized by

$$\hat{\beta} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}\,.$$

We can then estimate the covariance matrix by

$$\hat{\mathrm{Cov}}(\hat{\beta}) = (\mathbf{H}^T\mathbf{H})^{-1}\hat{\sigma}^2\,, \quad \text{where} \quad \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} \big(y - f(x_i)\big)^2\,.$$

For $h(x) = (h_1(x), \ldots, h_K(x))$, the standard error can then be estimated by

$$\hat{\mathrm{se}}(x) = [h(x)^T(\mathbf{H}^T\mathbf{H})^{-1}h(x)]^{\frac{1}{2}}\,\hat{\sigma}\,.$$

Bootstrapping

Once we have an estimate for the covariance and the standard error,

$$\hat{\mathrm{se}}(x) = [h(x)^T (\mathbf{H}^T \mathbf{H})^{-1} h(x)]^{\frac{1}{2}} \hat{\sigma}\,.$$

we can write the confidence confidence band as

$$\hat{f}(x) \pm 1.96 \times \hat{\mathrm{se}}(x)\,.$$

This confidence band gives 95% us an estimate of the true values of any point, given the assumption that the residual noise is Gaussian.

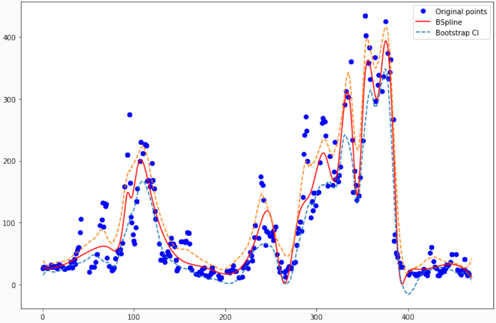
Bootstrapping

To bootstrap the confidence band, we sample B new training sets $\mathcal{T}_b$ from a uniform distribution on the training set $\mathcal{T}$. In general, we expect around 36.8% (or $1/e$) of the data to be contained in each bootstrap set $\mathcal{T}_b$.

For each $b = 1, \ldots, B$, we fit a regressor (or smoother) $\hat{f}_b(x)$ to the dataset $\mathcal{T}_b$. Then, at each point x, we generate a vector $(\hat{f}_1(x), \ldots, \hat{f}_B(x))$.

To compute the 95% confidence interval pointwise, we just take the value corresponding to the top 2.5% and the bottom 2.5% data points.

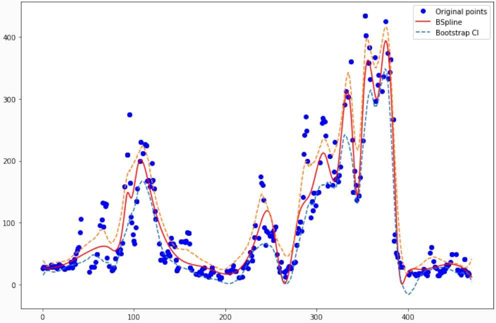
Bootstrapping



Bootstrapping by sampling with replacement is called the **nonparametric bootstrap**. The **parametric bootstrap** samples from a fit model, with Gaussian noise added and standard deviation given by the fit:

$$y_i^* = \hat{f}(x_i) + \epsilon_i\,, \qquad \epsilon_i \sim N(0, \hat{\sigma}^2)\,.$$

Bootstrapping



If the errors are actually Gaussian, the confidence interval computed via parametric bootstrap is equal to the least squares band as $B \to \infty$. This gives us a handy way to calculate the least squares band if the basis functions are unknown.

Parametric Bootstrapping and Maximum Likelihood

Interpretation of Bootstrapping

Algorithmically, the bootstrap is straight forward, but to understand what's being done statistically we need to discuss **frequentist** and **Bayesian** inference.

Roughly, frequentist inference draws conclusions from data by requiring that the correct conclusion should be drawn with a given (high) probability, from set of repeated drawings. Unknown parameters are assumed to have a fixed but unknown value and are not treated as random variables. They are then fit as exactly as possible, with confidence intervals measuring the uncertainty.

Bayesian inference allows parameters to be treated as coming from probability distributions themselves. The result of Bayesian inference is usually a probability distribution on the space of parameters.

Bootstrapping and Maximum Likelihood

The parametric bootstrapped confidence interval agrees with the least squares confidence interval because we have assumed additive Gaussian error. In general, the parametric bootstrap agrees with maximum likelihood and the frequentist notion of fitting.

Maximum likelihood starts by specifying a probability density $g_\theta(z)$ for our observations. We fit the density to data by maximizing the likelihood that i.i.d. sample $\{z_i\}$ are drawn from $g_\theta(z)$:

$$L(\theta; \mathbf{Z}) = \prod_{i=1}^{N} g_\theta(z_i)\,.$$

Maximizing $L(\theta; \mathbf{Z})$ is equivalent to maximizing the **log-likelihood**

$$\ell(\theta; \mathbf{Z}) = \sum_{i=1}^{N} \ell(\theta; z_i) = \sum_{i=1}^{N} \log g_\theta(z_i)\,.$$

Bootstrapping and Maximum Likelihood

The **score** is the derivative of the log-likelihood $\frac{\partial \ell}{\partial \theta}$, and is 0 at the maximizing value $\hat{\theta}$. The second moment of the score (the **Fisher information**))

$$\mathbf{i}(\theta) = E\left[\frac{\partial \ell}{\partial \theta}\middle|\theta\right]^2 = \int \left(\frac{\partial \ell}{\partial \theta}\right)^2 g_\theta(z)\, dz$$

can be rewritten by integration by parts as

$$\mathbf{i}(\theta) = -E\left[\frac{\partial^2 \ell}{\partial \theta \partial \theta^T}\middle|\theta\right] .$$

The **information matrix** then is defined as

$$\mathbf{I}(\theta) = -\sum_{i=1}^{N} \frac{\partial^2 \ell(\theta; z_i)}{\partial \theta \partial \theta^T} .$$

Bootstrapping and Maximum Likelihood

The sampling distribution of the maximum likelihood estimator can be shown to limit to a normal distribution around the true maximum θ_*

$$\hat{\theta} \rightarrow N(\theta_*, \mathbf{i}(\theta_*)^{-1})\,.$$

This suggest that we can approximate the distribution of $\hat{\theta}$ by

$$N(\hat{\theta}, \mathbf{I}(\hat{\theta})^{-1})\,.$$

Bootstrapping and Maximum Likelihood

Lets compute the information for the bootstrapped model with Gaussian noise ϵ_i. For Gaussian noise,

$$g_\theta(z) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(z-\mu)^2/\sigma^2} ,$$

and so the log-likelihood is

$$\ell(\theta) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - h(x_i)^T\beta)^2 .$$

The maximum likelihood occurs when $\frac{\partial \ell}{\partial \beta} = 0$, or when

$$\hat{\beta} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y} , \quad \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}(x_i))^2 .$$

The information matrix then is $\mathbf{I}(\theta) = (\mathbf{H}^T\mathbf{H})/\hat{\sigma}^2$, and so the variance of $\hat{\theta} \sim \mathcal{N}(\theta_*, \mathbf{I}(\theta)^{-1})$ agrees with the least squares estimate.

Bootstrapping and Maximum Likelihood

Recall that as a linear model the parametric bootstrap fits

$$\hat{f} = h^T(x)(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}\,.$$

and then estimates the covariance by

$$\hat{\mathrm{Cov}} = h^T(x)(\mathbf{H}^T\mathbf{H})^{-1}h(x)\hat{\sigma}^2\,.$$

But this just tells us that we're drawing $\theta \sim \mathcal{N}(\theta_*, \mathbf{I}(\theta)^{-1})$, exactly as previous slide showed for maximum likelihood. So one way that we can interpret the parametric bootstrap is as an estimate of the parameter variance we would derive from maximum likelihood if the underlying noise was in fact Gaussian.

Nonparametric Bootstraping and Bayesian Decision Making

Bootstrapping and Maximum Likelihood

In the Bayesian approach to decision making, we start by specifying a sampling model $P(\mathbf{Z}|\theta)$ (probability of sampling $\mathbf{Z} = (\mathbf{x}, \mathbf{y})$ given the parameters θ). Additionally, assume θ are drawn from a distribution $\mathrm{P}(\theta)$.

We start by sampling from $\mathrm{P}(\theta)$ some initial value of θ. We then look at our data and update our parameter estimates based on Bayes rule:

$$\mathrm{P}(\theta|\mathbf{Z}) = \frac{\mathrm{P}(\mathbf{Z}|\theta)\mathrm{P}(\theta)}{\int \mathrm{P}(\mathbf{Z}|\theta) \cdot \mathrm{P}(\theta)\, d\theta}\,.$$

In Bayesian inference, $P(\theta)$ is called the **prior** distribution and $P(\theta|\mathbf{Z})$ is called the posterior distribution. Predictions on new data based on old can then be made via

$$\mathrm{P}(z_{new}|\mathbf{Z}) = \int \mathrm{P}(z_{new}|\theta)\mathrm{P}(\mathbf{Z}|\theta)\,.$$

Bootstrapping and Maximum Likelihood

Lets work through this slowly with our spline example: Start by assuming $\beta \sim N(0, \tau\Sigma)$, where $\tau \in \mathbb{R}^+$ and Σ is estimated based on prior knowledge, or is set to I.Then

$$P(\beta) = \frac{1}{\sqrt{2\pi\tau|\Sigma|}} e^{-\frac{1}{2\tau}\beta^T\Sigma^{-1}\beta}, \qquad P(\mathbf{Z}|\theta) = \frac{1}{\sqrt{2\pi}\hat{\sigma}} e^{-\frac{1}{2}(y-h^T(x)\beta)^2/\hat{\sigma}^2}$$

To estimate $\mathrm{E}(\theta|\mathbf{Z})$, notice the denominator $\int \mathrm{P}(\mathbf{Z}|\theta) \cdot \mathrm{P}(\theta)\, d\theta$ contains no dependence on theta and the numerator is Gaussian, so the expectation value $\mathrm{P}(\mathbf{Z}|\theta)\mathrm{P}(\theta)$ will be at the minimum of

$$\mathrm{E}\big[\frac{1}{\tau}\beta^T\Sigma^{-1}\beta + (y - h^T(x)\beta)^2/\hat{\sigma}^2\big] \approx \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{N}(y - h^T(x)\beta)^2 + \frac{1}{\tau}\beta^T\Sigma^{-1}\beta\,.$$

This is just a form of ridge regression.

Bootstrapping and Maximum Likelihood

The posterior distribution for β can now be computed to be

$$\mathsf{E}(\theta|\mathbf{Z}) = \left(\mathbf{H}^{T}\mathbf{H} + \frac{\hat{\sigma}^2}{\tau}\Sigma^{-1}\right)^{-1}\mathbf{H}^{T}\mathbf{y}\,,$$

and

$$\mathsf{Cov}(\theta|\mathbf{Z}) = \left(\mathbf{H}^{T}\mathbf{H} + \frac{\hat{\sigma}^2}{\tau}\Sigma^{-1}\right)^{-1}\hat{\sigma}^2\,.$$

Notice that as $\tau \to \infty$, this becomes a constant prior and in turn reduces to the linear regression fit. We again see that Bayesian reasoning leads to ridge regression, and in turn that another way to understand the λ of ridge regression is as an inverse variance in the parameters.

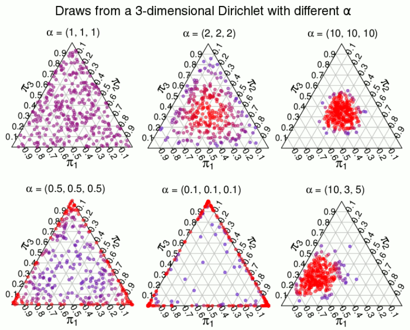
Bootstrapping and Maximum Likelihood

Lets return to the nonparameteric bootstrap, and show how it fits into Bayesian inference. Let $\mathcal{Y}_U$ be the set of unique y_i values, and defined weights w_j, $j = 1, \ldots, L$, by their frequency. For y continuous, it is almost certain that $\mathcal{Y}_U = \{y_i\}$ and $w_j = \frac{1}{N}$.

We then estimate a prior distribution on the space of label weights w_j by a Dirichlet distribution $w_i \sim \mathrm{Di}_L(1, \ldots, 1)$. We then sample the data for empirical probabilities $\hat{w}_i$, and update the prior distribution to a posterior distribution.

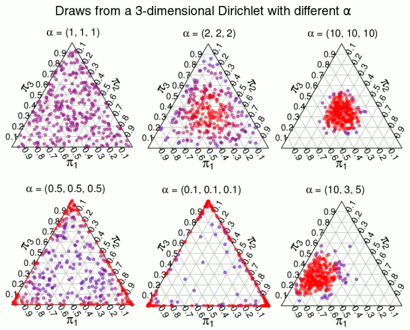
Finally, we will compare the result to the mutlinomial distribution from which the bootstrap draws $\hat{w}_b \sim \mathrm{Mult}(N, \hat{w})$.

Bootstrapping and Maximum Likelihood



The Dirichlet distribution $\mathrm{Di}_L(\alpha_1, \alpha_2, \alpha_3) \propto \prod_i w_i^{\alpha_i}$ is pictured above and gives a natural candidates for the weights. It interpolates probabilities between L different features in a smooth way.

Bootstrapping and Maximum Likelihood

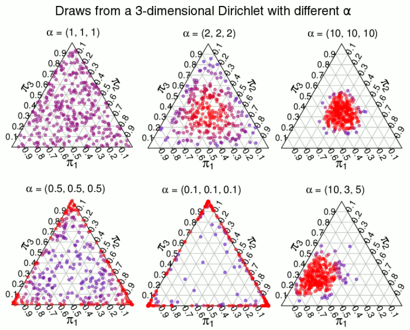


After sampling the weights, the posterior density is

$$\mathrm{Di}_L(a + N\hat{w}) \to \mathrm{Di}_L(N\hat{w}) \, ,$$

where we have taken the low information limit $a \to 0$. The multinomial closely matches $\mathrm{Di}_L(N\hat{w})$, only the covariance matrix differs slightly.

Bootstrapping and Maximum Likelihood



Practically, this means the non-parametric bootstrap is estimating the posterior distribution of a prior Dirichlet density over the labels y_i. The non-parametric bootstrap has been loving characterized as "the poor mans Bayesian confidence interval."

Bagging and Bumping

Bagging and Bumping

We will now go to the minilab to discuss **bagging** and **bumping**.

Bagging, or Bootstrap AGGragatING is the process of repeated sampling the training data, fitting a series of classifiers, and then taking their average, and has been shown empirically to lead not only to better results but to results with far less variance.

Bumping is like bagging, although we just take the best fit from our bootstrapped fits.

Reference

The main reference for this lecture is Chapter ESLII Chapter 8.

For a deeper perspective on the Bayesianism, the nonparametric bootstrap, and Baysean bootstrapping see Rasmus Bååth's blog `http://www.sumsar.net/blog/2015/04/the-non-parametric-bootstrap-as-a-bayesian-model/`

For more information on bootstrapping and confidence intervals see `https://www.ethz.ch/content/dam/ethz/special-interest/math/statistics/sfs/Education/Advanced%20Studies%20in%20Applied%20Statistics/course-material-1719/Nonparametric%20Methods/lecture_2up.pdf`.