Bayesian Parameter Estimation

The problem of hypothesis testing can be thought of as estimating some discrete parameter i that determines which hypothesis H_0 or H_1 is true

- In Bayesian / minimax formulation, this is some random variable
- In Neyman-Pearson formulation, this is deterministic but unknown

We typically denote a parameter that we are estimating as $\mathbf{x}$ and $\mathbf{y}$ as the measurements we get

From Bayes Rule:

$$p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) = \frac{p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})p_{\mathbf{x}}(\mathbf{x})}{\int p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x'})p_{\mathbf{x}}(\mathbf{x'})d\mathbf{x'}}$$

Bayes Risk for Estimation

We use $\hat{\mathbf{x}}(\mathbf{y})$ to denote our estimate of $\mathbf{x}$ based on our observation $\mathbf{y}$

We need some "measure of goodness" to denote how good this estimator is

- We choose some scalar valued cost function $C(\mathbf{a}, \hat{\mathbf{a}})$
- We want our estimator to minimize the expected value of this, which is known as the **Bayes risk**

$$\mathbb{E}\left[C(\mathbf{x}, f(\mathbf{y}))\right]$$

  - This expectation is jointly over both $\mathbf{x}$ and $\mathbf{y}$
- Similar to the Bayes hypothesis testing case, this can be solved for by minimizing for each different value $\mathbf{y}$ separately
  - Our goal is then to, for each $\mathbf{y}$, find the $\mathbf{a}$ that minimizes

$$\int_{-\infty}^{\infty} C(\mathbf{x}, \mathbf{a})p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})d\mathbf{x}$$

When we can break down $C(\mathbf{x}, \mathbf{a}) = \sum_i C_i(x_i, a_i)$, then we can solve for each component separately

- For now, we will focus on just the scalar case for various cost functions

Minimum Absolute Error (MAE)

In MAE, the cost function is $C(a, \hat{a}) = |a - \hat{a}|$

> **Claim**
>
> The MAE estimate is the median of the posterior belief, that is, the value of x that that $\text{cdf}(x) = 1/2$

Sketch:

- Let the estimate be a
- We can break the absolute value into $x - a$ and $a - x$ and break the expression into two integrals with one going from $-\infty$ up to a and the other going from a to ∞
- Differentiating and setting to 0 yields that a must be positioned such that the densities on both sides are equal

We can generalize this to vectors via the L^1 norm, which is componentwise absolute values

Maximum A Posteriori (MAP)

Consider the minimum uniform cost (MUC) criterion with the cost function:

$$C(a, \hat{a}) = \begin{cases} 1 & |a - \hat{a}| > \varepsilon \\ 0 & \text{otherwise} \end{cases}$$

Claim

In the limit as $\varepsilon \to 0$, the MUC estimate is the mode of the posterior, that is, the argmax of the posterior

This is known as the MAP estimate

Sketch

- Choosing the mode makes it so that the most density is concentrated in interval around a
- Any other choice would have less density in the neighborhood and be worse off

Least Square Estimation

The most popular estimator is based on a quadratic cost criterion:
$$C(a, \hat{a}) =

- \hat{a}

^2$$

This is known as the Bayes least-squares (BLE) estimator

- Also known as minimum mean-square-error (MMSE) estimator

Claim

The BLS estimator is the mean of the posterior distribution

$$\mathbb{E}\left[x | y = y\right] = \int x p_{x|y}(x|y) dx$$

When $\mathbf{x}$ is a vector, this can be decomposed into components, so we can just minimize the error of each component separately

Properties

We write the error term $e(x, y) = \hat{x}(y) - x$

The **global bias** of the estimator is:

$$b = \mathbb{E}\left[e(x, y)\right]$$

The error covariance matrix is:

$$\Lambda_e = \mathbb{E}\left[(e(x, y) - b)(e(x, y) - b)^T\right]$$

The error correlation matrix is:

$$\mathbb{E}\left[ee^T\right] = \Lambda_e + bb^T$$

- The MSE is just the trace of this matrix

Claim

The BLS estimator is unbiased

This also means that the error correlation matrix is just the error covariance matrix

Claim

The error covariance matrix is just the expected covariance of the posterior belief:

$$\Lambda_{BLS} = \mathbb{E}\left[\Lambda_{x|y}(y)\right]$$

BLS Orthogonality

An estimator $\hat{x}$ is the BLS estimator iff the estimation error is orthogonal to any vector-valued function g of the data:

$$\mathbb{E}\left[(\hat{x}(y) - x)g(y)^T\right] = 0$$

The idea is the error should be uncorrelated with any function of the data we construct

- Therefore, we cannot transform the data to further reduce the error in our estimate