Lecture: Principles of Statistical Model Selection

Radu Balan

March 4, 2019

Problems for today

Model Selection based on statistical principles:

1. KL divergence as "distance" between models
2. Akaike Information Criterion (AIC)
3. Other criteria: BIC and MDL

Statistical Estimation of Model Parameters

Models and Likelihoods

Assume we perform a measurement $x \in \mathbb{R}^n$ of a random variable X. For X we assume a family of models that explain the measurement via a probability distribution function $p(x;\theta)$ parametrized by $\theta \in \Theta$.
The goal of this lecture is to find the "Most Likely" model that explains the measurement.
Approach: Assume the "true" distribution of data X is given by $p_X(x)$.
Then a statistically principled way of estimating the parameter θ is by minimizing a "distance" $D(p_X(x), p(x;\theta))$ between the two distributions over parameter θ:

$$\hat{\theta} = argmin_\theta D(p_X(x); p(x;\theta))$$

Kullbeck-Leibler Divergence

How to measure how far apart are two probability distribution functions

One choice for "distance" between probability distribution functions: the Kullback-Leibler divergence.

Assume $p, q : \mathbb{R} \to \mathbb{R}$ are two probability distributions functions, i.e., $p(x), q(x) \geq 0$ and $\int_{-\infty}^{\infty} p(x)dx = \int_{-\infty}^{\infty} q(x)dx = 1$.

Definition. The *Kullback-Leibler divergence* (or *KL "distance"*) between p and q denoted by $D(p||q)$ or $KL(p||q)$ is given by

$$D(p||q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx =: \mathbb{E}_{X \sim p} \left[\log \frac{p(X)}{q(X)} \right]$$

Kullbeck-Leibler Divergence
Properties

While not a distance between two pdf's (it is not symmetric, nor satisfy triangle inequality), the KL divergence satisfies:

Proposition

Assume p, q are two probability distribution functions. Then:

1. $D(p||q) \geq 0$
2. $D(p||q) = 0$ *if and only if* $p = q$.

Why:

1. Since the logarithm is concave $t\,log(r_1) + (1-t)\,log(r_2) \leq log(tr_1 + (1-t)r_2)$. By a limiting argument:

$$\int_{-\infty}^{\infty} p(x)\log(r(x))dx \leq \log\left(\int_{-\infty}^{\infty} p(x)r(x)dx\right)$$

(known as Jensen's inequality)

Kullbeck-Leibler Divergence
Properties - cont'ed

For $r(x) = \frac{q(x)}{p(x)}$ we obtain:

$$-D(p||q) = \int_{-\infty}^{\infty} p(x) \log \frac{q(x)}{p(x)} dx \leq \log \left(\int_{-\infty}^{\infty} p(x) \frac{q(x)}{p(x)} dx \right) = \log(1) = 0.$$

Hence $D(p||q) \geq 0$.
2. The proof also shows when equality is achieved: $D(p||q) = 0$ only if equality in Jansen's inequality. Since *log* is a strictly concave function, equality is achieved only when the argument of *log*() is constant on its support. Hence $p = q$.
Note: Similar formula applies for vector-valued random variables:

$$D(p||q) = \int_{\mathbb{R}^n} p(x) \log \frac{p(x)}{q(x)} d^n x$$

Maximum Likelihood Estimation
MLE

The *frequentist* approach to estimating parameter (vector) $\theta \in \mathbb{R}^d$:

$$\hat{\theta} = argmin_\theta D(p_X || p(\cdot;\theta))$$

Note:

$$D(p_X || p(\cdot;\theta)) = \int p(x) log(p(x)) dx - \mathbb{E}[log(p(X;\theta))]$$

Hence the minimizer above is the maximizer in:

$$\hat{\theta} = argmax_\theta \mathbb{E}[log(p(X;\theta))] = \int_{\mathbb{R}^n} p(x) \log(p(x;\theta)) dx$$

Maximum Likelihood Estimation
MLE - 2

Assume we are given a set of measurements $\{x_1, \cdots, x_T\}$ each an independent realization of the same random (vector) variable X. Then we approximate the expectation with respect to the "true" unknown distribution p_X with the sample mean:

$$\mathbb{E}[log(p(X;\theta))] \approx \frac{1}{T}\sum_{t=1}^{T} log(p(x_t;\theta))$$

We obtain the "most likely" explanation of the measurements is given by the model whose parameter vector θ is:

$$\hat{\theta}_{MLE} = argmax_\theta \sum_{t=1}^{T} log(p(x_t;\theta))$$

This estimator is called the *Maximum Likelihood Estimator* (MLE) of parameter θ. The functions $p(X;\theta)$ are called "likelihoods".

LS Estimator as MLE for AWGN

Consider the case of data points:

$$Y_t = A(X_t - z) + \nu_t \quad , \quad \nu_t \sim \mathbb{N}(0, \sigma^2 I_d) \ , \ 1 \leq t \leq n$$

where the parameters are $\theta = (A, z)$ and measurements (Y, X). The likelihood is then:

$$p(Y, X; A, z) = \frac{1}{(\sqrt{2\pi}\sigma)^{dn}} exp\left(-\frac{1}{2\sigma^2}\|Y - AX + Az1^T\|_F^2\right)$$

It follows the MLE estimator for θ is the one that minimizes:

$$(\hat{A}, \hat{z}) = argmin_{A,z}\|Y - A(X - z1^T)\|_F^2$$

Hence the MLE for Additive White Gaussian Noise (AWGN) model reduces to the Least Squares Estimator (LSE).

Why not estimating the number of parameters through MLE?

You might be tempted to include the number of parameters as an additional parameter (in θ) and estimate it accordingly.
Specifically, consider the following natural succession of models, each defining the matrix A:

$$M_1 \subset M_2 \subset M_3$$

where:

$$M_1 = \mathbb{R}^+ \cdot I_d = \{aI_d, a > 0\} \ , \quad M_2 = \mathbb{R}^+ \cdot SO(d) = \{aQ \ , \ Q \in SO(d)\}$$

$$M_3 = GL(d, \mathbb{R}) = \{A : det(A) \neq 0\}$$

Why not estimating the number of parameters through MLE?

You might be tempted to include the number of parameters as an additional parameter (in θ) and estimate it accordingly.
Specifically, consider the following natural succession of models, each defining the matrix A:

$$M_1 \subset M_2 \subset M_3$$

where:

$$M_1 = \mathbb{R}^+ \cdot I_d = \{aI_d, a > 0\} \quad , \quad M_2 = \mathbb{R}^+ \cdot SO(d) = \{aQ \ , \ Q \in SO(d)\}$$
$$M_3 = GL(d, \mathbb{R}) = \{A : det(A) \neq 0\}$$

Due to nestedness of these models, the more complex models always provide a better fit to data (i.e., smaller residual errors). However this does not imply a better model!
Sometime this is referred to as the "data overfitting".

How to fix the problem?
Akaike Principle

Akaike introduces a penality term to penalize model complexity. Specifically, let $p(Data;\theta)$ denote the likelihood of a model parametrized by a D-vector θ, Hence D represents the number of parameters. Let

$$J(\hat{\theta};D) = min_\theta\left[-logp(Data;\theta)\right]$$

denote the minimum negative log-likelihood (equal to the negative maximum log-likelihood). Then Akaike adds a penality term equal to the number of parameters:

$$minimize_D \;\; J(\hat{\theta};D) + D$$

The rational for this choice is the fact that MLE of parameter θ produces a random variable $\hat{\theta}_{MLE}$ which, according to the central limit theorem, asymptotically is distributed like a normal random variable (i.e. Gaussian) centered not at the true value θ but biased by D.

Akaike Information Criterion

The Akaike Information Criterion (AIC) is used not only to estimate the model parameters, but rather to select between models:

$$AIC = minimize_D \left[-maximize_\theta log p(Data;\theta) + D\right]$$

The first term reflects the fact that more complex models always provide a better fit to Measured Data. However the second term represents a penalty for using more "complicated" models. It increases with model complexity.

Other Information Theoretic Criteria

The Bayesian Information Criterion (BIC) and the Minimum Description Length (MDL)

Here is a summary of three Information Theoretic criteria for model selection:

$$AIC = minimize_D\left[-max_\theta \, log \, p(Data;\theta) + D\right]$$

$$BIC = minimize_D\left[-max_\theta \, log \, p(Data;\theta) + D\frac{log(T)}{2}\right]$$

$$MDL = minimize_D\left[-max_\theta \, log \, p(Data;\theta) + CodingLength(Model(\theta, D))\right]$$

References