Statistical Processes Lecture 3

Physics 129AL

Zihang Wang
10/18/2023

Data Analysis with Bayesian

Let's say you have N measurements (N degrees of freedom) from an experiment. Can you reduce the degrees of freedom to M parameters (M<<N)? Of course, the DOF reduction is not perfect due to the random (or systematic) errors that associate with the data.

- We are given N number of data measurements (x_i, y_i)
- Each measurement comes with an error estimate σ_i
- We have a parametrized model for the data $y = y(x_i)$
- We think the error probability is Gaussian and the measurements are uncorrelated:

$$p(y_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{-(y(x_i)-y_i)^2}{2\sigma_i^2}}$$

$$p(\vec{y}) = \prod_i p(y_i)$$

Data Analysis with Bayesian

We can parametrize the model in terms of M free parameters

$y(x_i|a_1,a_2,a_3,\ldots,a_M)$

Bayesian formalism gives us the full posterior information on the parameters of the model

$$p(\vec{y}|\vec{a}) = \prod_i p(y_i|\vec{a}) = \mathcal{L}(\vec{a})$$

$$p(a_1,\ldots,a_M|\vec{y}) = \frac{\prod_i p(y_i|\vec{a})p(\vec{a})}{p(y_i)}$$

We can assume a flat prior $p(a_1,a_2,a_3,\ldots,a_M) = \text{constant}$

Maximum Likelihood with Gaussian Errors

- Instead of the full posterior we can ask what is the best fit value of parameters $a_1, a_2, a_3, \ldots, a_M$
- We can define this in different ways: mean, median, mode
- Choosing the mode (peak posterior or peak likelihood) means we want to maximize the likelihood: maximum likelihood estimator (or MAP for non-uniform prior)

$$\text{MLE}: \quad \frac{\partial \mathcal{L}}{\partial \vec{a}} = 0 \quad \text{or} \quad \frac{\partial \ln \mathcal{L}}{\partial \vec{a}} = 0$$

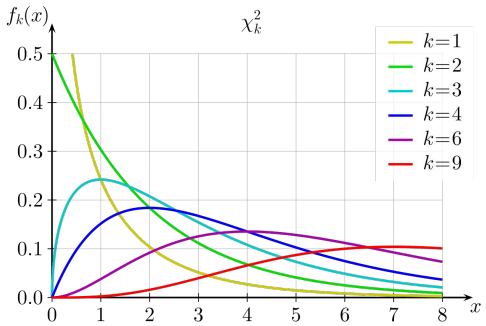
Maximum Likelihood with Gaussian

$$-2\ln\mathcal{L} = \underbrace{\sum_i \left\{ \frac{(y_i - y(x_i|a_1, ..., a_M))^2}{\sigma_i^2} + \ln\sigma_i \right\}}_{\chi^2}$$

Since σ_i does not depend on a_i, MLE means minimizing χ^2 wrt a_k

$$\frac{\partial\chi^2}{\partial a_k} = 0 \quad \rightarrow \quad \sum_i \frac{y_i - y(x_i)}{\sigma_i^2} \frac{\partial y(x_i)}{\partial a_k} = 0$$

Chi-square Distribution with degrees of freedom k.



Linear Regression: Fitting data to a straight line

Linear Regression $y(x) = y(x; a, b) = a + bx$

$$-2\log(\mathcal{L}) = -2\log\left(\prod_i p(y_i|a,b)\right)$$
$$= \sum_i \frac{(y_i - y(x_i|a,b))^2}{\sigma_i^2} + \log(\sigma_i) = \chi^2(a,b) + \sum_i \log(\sigma_i)$$

where $p(y_i|a,b)$ is the likelihood function. We then can minimize the χ^2 with respect to the controlling parameters a, b.

Minimize χ^2:

$$0 = \frac{\partial \chi^2}{\partial a} = -2\sum_{i=1}^{N} \frac{y_i - a - bx_i}{\sigma_i^2}$$

$$0 = \frac{\partial \chi^2}{\partial b} = -2\sum_{i=1}^{N} \frac{x_i(y_i - a - bx_i)}{\sigma_i^2}$$

Linear Regression: Fitting data to a straight line

Define:
$$S \equiv \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \quad S_x \equiv \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \quad S_y \equiv \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2}$$
$$S_{xx} \equiv \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} \quad S_{xy} \equiv \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2}$$

$$0 = \frac{\partial \chi^2}{\partial a} = -2 \sum_{i=1}^{N} \frac{y_i - a - b x_i}{\sigma_i^2}$$
$$0 = \frac{\partial \chi^2}{\partial b} = -2 \sum_{i=1}^{N} \frac{x_i (y_i - a - b x_i)}{\sigma_i^2}$$

$$\longrightarrow \quad aS + bS_x = S_y$$
$$aS_x + bS_{xx} = S_{xy}$$

Matrix Form:
$$\begin{pmatrix} S & S_x \\ S_x & S_{xx} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} S_y \\ S_{xy} \end{pmatrix}$$

Solve this with linear algebra

Linear Regression: Fitting data to a straight line

$$\begin{pmatrix} S & S_x \\ S_x & S_{xx} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} S_y \\ S_{xy} \end{pmatrix}$$

$$C^{-1} = \begin{pmatrix} S & S_x \\ S_x & S_{xx} \end{pmatrix}$$

$$C = \frac{1}{\Delta} \begin{pmatrix} S_{xx} & -S_x \\ -S_x & S \end{pmatrix}$$

Solution: Define $\Delta \equiv SS_{xx} - (S_x)^2$

$$\hat{a} = \frac{S_{xx}S_y - S_x S_{xy}}{\Delta}$$

$$\hat{b} = \frac{SS_{xy} - S_x S_y}{\Delta}$$

This gives best fit $\hat{a}$ & $\hat{b}$

This is also known as the least square method or maximum likelihood method.

Linear Regression: Fitting data to a straight line

Since we assume a uniform prior, the posterior proportional to likelihood function in Bayesian formalism.

The fixed point solution is given by $\hat{a}, \hat{b}$. What about the second order variation of hte likelihood function? With Taylor expansion near the fixed point, we have the Hessian,

$$(-2\log(\mathcal{L}(a,b))) = \left(-2\log(\mathcal{L}(\hat{a},\hat{b}))\right)$$
$$= \frac{1}{2}\left((a-\hat{a})^2\frac{\partial^2}{\partial^2 a} + (a-\hat{a})(b-\hat{b})\frac{\partial}{\partial a}\frac{\partial}{\partial b} + (b-\hat{b})(a-\hat{a})\frac{\partial}{\partial b}\frac{\partial}{\partial a} + (b-\hat{b})^2\frac{\partial^2}{\partial^2 b}\right)\Big|_{a,b}(-2\log(\mathcal{L}))$$

Linear Regression: Fitting data to a straight line

We then define the correlation matrix,

$$-\frac{\partial^2 \ln \mathcal{L}}{\partial x_i \partial x_j} \equiv C_{ij}^{-1}$$

And the inverse is called precision matrix. This is called Gaussian posterior approximation: we are dropping terms beyond 2 nd order.

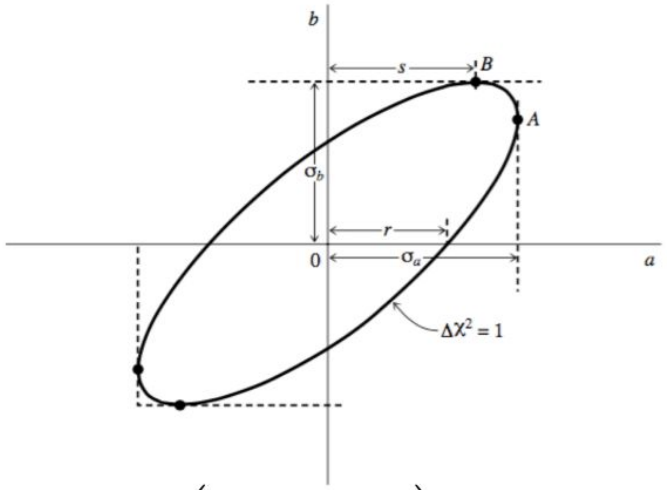
Linear Regression: Fitting data to a straight line

$$-2 \cdot \ln\mathcal{L} = \chi^2$$

$$\frac{\partial^2\chi^2}{\partial a^2} = 2\sum_i \frac{1}{\sigma_i^2} = 2S$$

$$\frac{\partial^2\chi^2}{\partial b^2} = 2\sum_i \frac{x_i^2}{\sigma_i^2} = 2S_{xx}$$

$$\frac{\partial^2\chi^2}{\partial a\partial b} = 2\sum_i \frac{x_i}{\sigma_i^2} = 2S_x$$



$$C^{-1} = \begin{pmatrix} S & S_x \\ S_x & S_{xx} \end{pmatrix} \qquad C = \frac{1}{\Delta}\begin{pmatrix} S_{xx} & -S_x \\ -S_x & S \end{pmatrix}$$

S^{-1} is error on a at a fixed b

Define $\Delta \equiv SS_{xx} - (S_x)^2$

$$\sigma_a^2 = S_{xx}/\Delta \qquad \sigma_b^2 = S/\Delta$$

Linear Regression: Fitting data to a straight line

- The posterior distribution $p(a,b|y_i)$ is described as a 2-d C^{-1} ellipse in (a,b) plane
- At any fixed value of a (or b) the posterior of b (or a) is a gaussian with variance $[C^{-1}{}_{bb(aa)}]^{-1}$
- If we want to know the error on b (or a) independent of a (or b) we need to marginalize over a (or b)
- This marginalization can be done analytically (completion of squares), and leads to $C_{bb(aa)}$ as the variance of b (or a)
- This will increase the error: $C_{bb(aa)} > [C^{-1}{}_{bb(aa)}]^{-1}$

Asymptotics theorems (Gaussian posterior)

- At a fixed number of parameters posteriors approach a multi-variate Gaussian in the large N limit (N: number of data points):

this is because the 2nd order Taylor expansion of ln L is more and more accurate in this limit, i.e. we can drop 3rd and higher order terms, by central limit theorem

- The marginalized means approach the true value and the variance approaches the Fisher matrix, defined as ensemble average of precision matrix $\langle C^{-1} \rangle$

- The likelihood dominates over the prior in large N limit

Asymptotics theorems (Gaussian posterior)

- There are caveats when this does not apply, e.g. when data are not informative about a parameter or some linear combination of them, when number of parameters M is comparable to N, when posteriors are improper or likelihoods are unbounded… Always exercise care!

- In practice the asymptotic limit is often not achieved for nonlinear models, i.e. we cannot linearize the model across the region of non-zero posterior: this is why we will use advanced Bayesian methods to evaluate the posteriors instead of Gaussian

Multivariate linear least squares

- We can generalize the model to a generic functional form

$$y_i = a_0 X_0(x_i) + a_1 X_1(x_i) + \ldots + a_{M-1} X_{M-1}(x_i)$$

- The problem is linear in a_j and can be nonlinear in x_i,

  e.g. $X_j(x_i)=x_i^j$

$$\chi^2 = \sum_{i=0}^{N-1} \left[\frac{y_i - \sum_{k=0}^{M-1} a_k X_k(x_i)}{\sigma_i} \right]^2$$

- We can define design matrix $A_{ij} = X_j(x_i)/\boldsymbol{\sigma}_i$ and

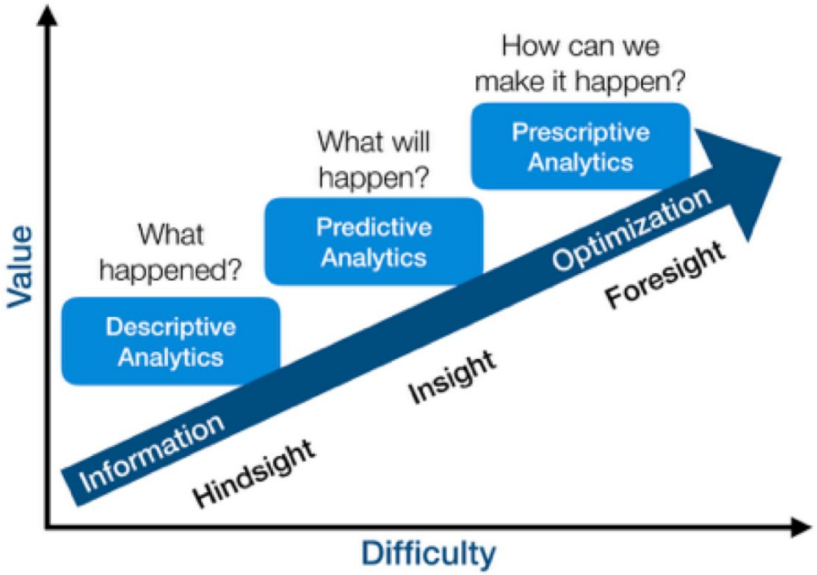
- $b_i = y_i/\boldsymbol{\sigma}_i$

$$\boxed{\chi^2 = |\mathbf{A} \cdot \mathbf{a} - \mathbf{b}|^2}$$

Learning from data

In physics, we usually use the data to validate existing analytical models for future model-based predictions. This is called **Data analysis**.

What happen if we do not know the analytical model for a given dataset? In this case, we need to ask the machine to summarize the data or make predictions. This is called **machine learning (optimizations).**



Chris Wiggins taxonomy, Gartner/Recht graph

Type of Machine Learning

Let’s say we have a dataset, but we do not have any existing model for it.

If we want to use the existing dataset to make future predictions without knowing the analytical model, this is called **supervised learning**.

If we want to know about what we can conclude from the dataset, this is called **unsupervised learning**.

If we want to maximize reward base on environmental response, this is called **reinforcement learning**.

Supervised machine learning procedure

- We have some data x and some labels y, such that Y=(x,y). We wish to find some model g(a) and some cost or loss function C(Y, g(a)) that we wish to minimize such that the model g(a) explains the data Y.
- E.g. Y=(x,y), C=χ^2
- G=$a_0X_0(x_i) + a_1X_1(x_i) + \ldots + a_{M-1}X_{M-1}(x_i)$

$$\chi^2 = \sum_{i=0}^{N-1} \left[\frac{y_i - \sum_{k=0}^{M-1} a_k X_k(x_i)}{\sigma_i} \right]^2$$

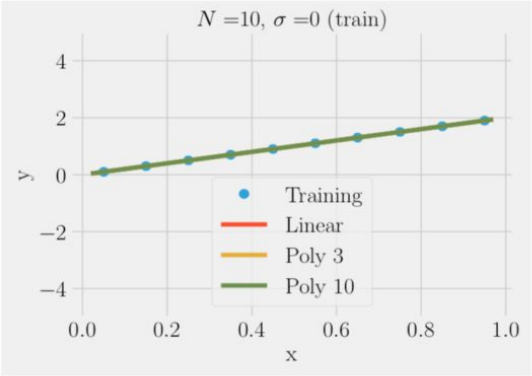
- In ML we divide data into training data Y_{train} (e.g. 90%) and test data Y_{test} (e.g. 10%)
- We fit model to the training data: the value of the minimum loss function at a_{min} is called in-sample error $E_{in}=C(Y_{train},g(a_{min}))$
- We test the results on test data, getting out of sample error $E_{out}=C(Y_{test},g(a_{min}))>E_{in}$
- This is called cross-validation technique

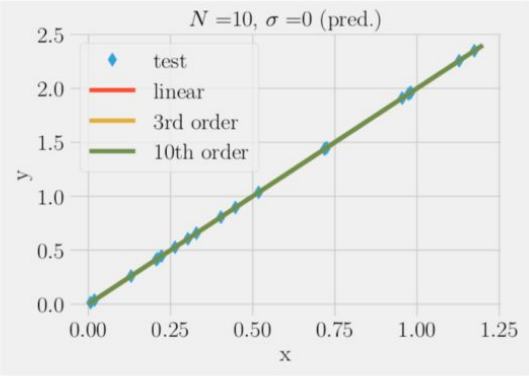
Summary:

- Data analysis: fitting existing data to a physics based model to obtain model parameters y. Parameters are fixed: we know physics up to parameter values. Parameter posteriors are the goal.
- ML: use model derived from existing data to predict regression or classification parameters y for new data.

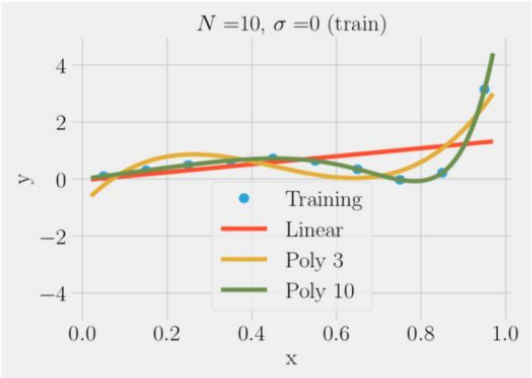
- We can fit the training data to a simple model or complex model
- In the absence of noise complex model (many fitting parameters a) always better
- In the presence of noise complex model often worse
- Note that parameters a have no meaning on their own, just means to reach the goal of predicting y

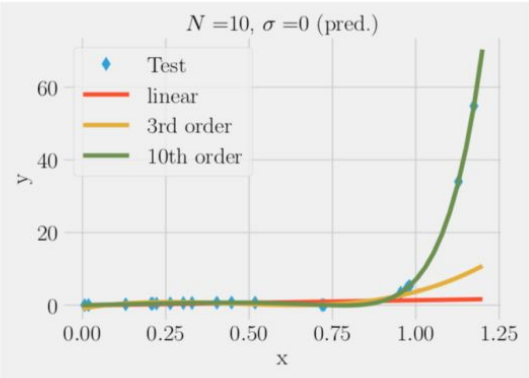
Zero Noise:





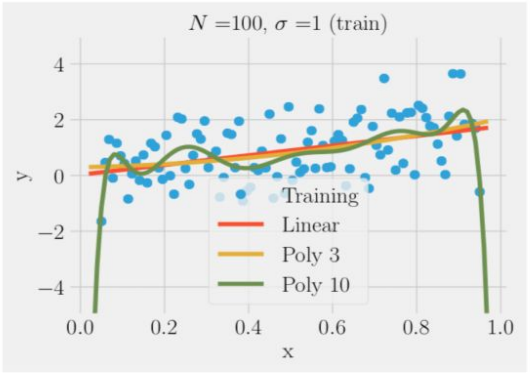
f(x)=2x, no noise

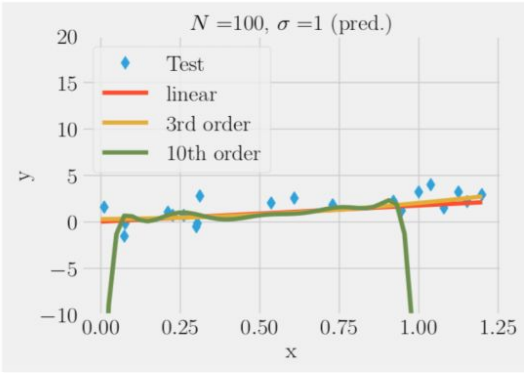




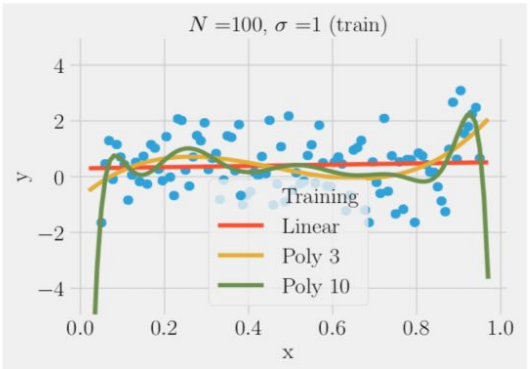
f(x)=2x-10x^5+15x^{10}

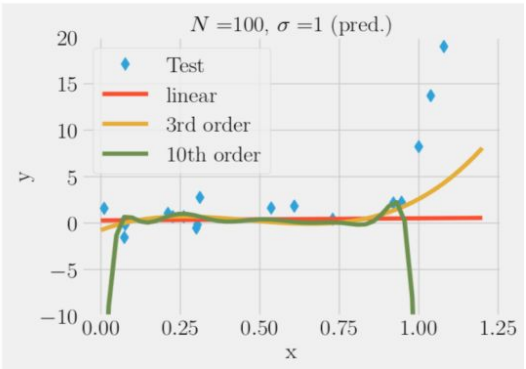
Non-zero Noise:





Over-fitting noise with too complex models (bias-variance trade-off)

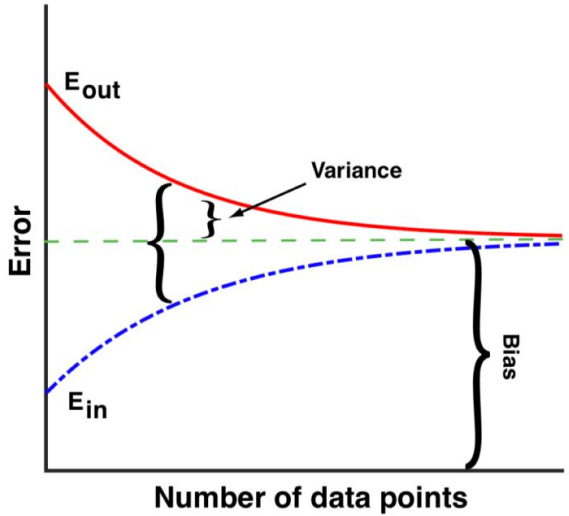




Trade off at fixed model complexity

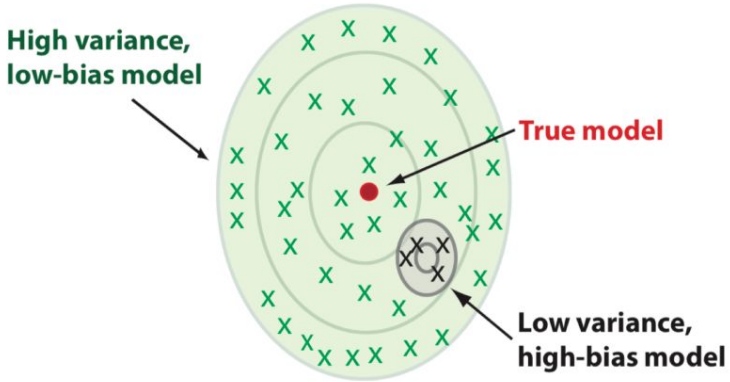
Small data size suffers from a large variance, and large data size suffers from model bias.

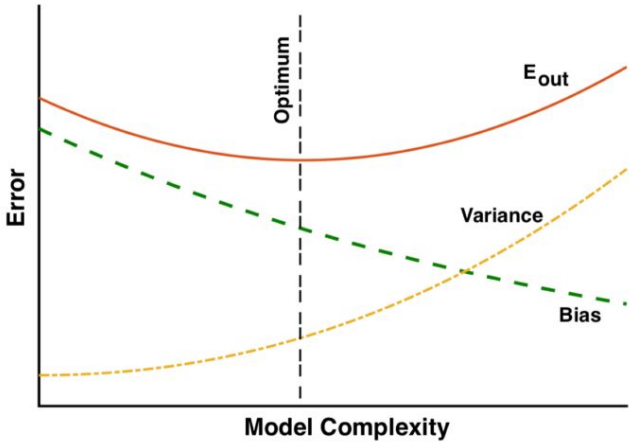
- We fit model to the training data: the value of the minimum loss function at a_{min} is called in-sample error $E_{in}=C(Y_{train},g(a_{min}))$
- We test the results on test data, getting out of sample error $E_{out}=C(Y_{test},g(a_{min}))>E_{in}$



Bias-variance trade-off vs complexity

Low complexity suffers from a large bias, and large complexity suffers from a large variance.





Acknowledgement

The slides are partially developed or inspired by Professor Uros Seljak at UC Berkeley. For more information, please visit the github page.

A short story: When I was doing my undergraduate, I took his class (the exact same one, linked below).

Uros Seljak (Berkeley)

https://phy151-ucb.github.io/seljak-phy151-fall-2018/#course-syllabus