

Decision Making: An Introduction

Song Liu (song.liu@bristol.ac.uk)

Twitter: @songandylilu

Office: Fry Building, GA.18

How to be a PhD student?

- Things to do with your supervisor:
 - Decide core working hours (e.g., 10am - 3pm), so you and your supervisor can easily reach each other via email or IM and start discussions during this time.
 - Schedule a fixed meeting time slot. The frequency depends on you and your supervisor's preference. Most common period is weekly or biweekly.
- Things to keep in mind:
 - Stick to your meeting schedule. Even if there is not much to report, it is still nice to update him/her regularly on your research.
 - Get up early! Working at night constantly causes mental health issues.
 - Go to seminars (in particular, statistics seminars.)

Prologue

- Unit Director: Dr. Song Liu (Office GA 18)
- Who am I?
 - A former MSc student in the University of Bristol, 14 years ago.
 - Went to Japan for my PhD and Postdoc.
 - Came back to work as a lecturer in Statistical Science
 - Why?
 - Homepage: <http://allmodelsarewrong.net>
- What do I do?
 - *intractable model inference, estimating statistical discrepancies, and their applications (such as Score Matching and GAN).*



Prologue

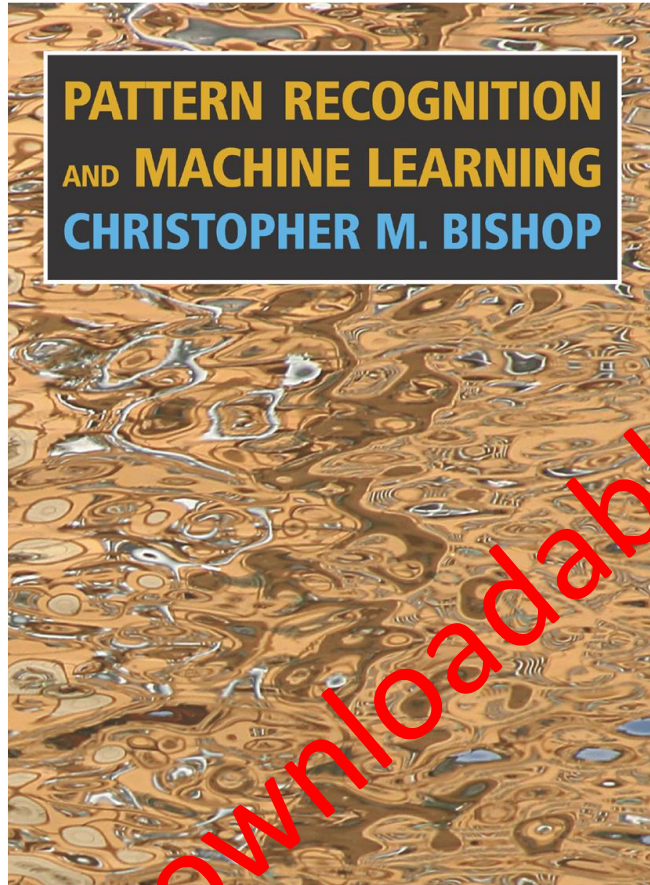
- **Two Classes (Lectures) + One Computing Lab. (Practice)**
 - Classes: Monday 12-1pm and Wednesday, 9-10am
 - Lab: Friday, 10am-12, 2 hours.
- **Assessment Plan (Read online document):**
 - **5** Personal portfolio (30%)
 - Summary of lectures, in your own words
 - Answers to Homework.
 - **2 Assessed coursework (40%)**
 - **Announcement: Wednesday after lecture, Week 5 and Week 9**
 - **Deadline: Friday 5pm, Week 5 and Week 9**
 - **1** SM1 + SC1 Group project (30%)

Prologue

- **Syllabus:**

- Introduction of Statistical Decision Making/Learning Week 1-2
 - 4 lectures
- Probability Theory Week 3
 - 2 lectures
- Linear Methods for Regression Week 4-5
 - 3 lectures
- Linear Methods for Classification Week 6-7
 - 3 lectures
- Probabilistic Graphical Model Week 8
 - 2 lectures
- Advanced Topics in Machine Learning (2 guest lectures) Week 9

Reference



This unit **roughly** follows
Chapter 1,2,3 and 4 of

Pattern Recognition and
Machine Learning

Christopher Bishop, 2006

Decision Making

- Many modern-day computational tasks are about making decisions or predictions.



- Decision making has been a great challenge of human society for a long time.

◀◀ A Look back ... in China



“Oracle Bones”

- Emperor has a **question**.
- **Write it down** on the bones of large animals and **toss** it to **flame**.
- **Cracks on bones** reveal “Gods’ will”.
- **Priest deciphers** the patterns of cracks and provides an answer.

◀◀ A Look back ... in Greece



Pythia

- Supplicant has a **question**.
- He/she travels to Delphi **asks Pythia**.
- **Pythia** inhales vapors at Temples of Apollo, speaks gibberish.
- Priest deciphers her gibberish and provides supplicant an answer.

Fast forward ►► ... Modern Era

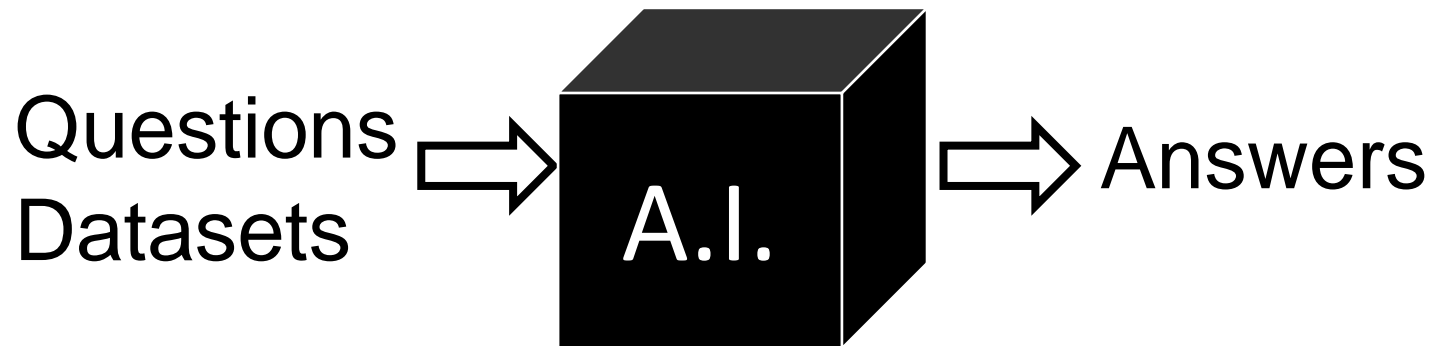
- No one believes in Pythia or Oracle bones anymore.
- However, the modern-day society faces another great challenge on decision making.



- Human cannot digest information fast enough and make rationalized decisions.

Fast forward ►► ... Modern Era

- Therefore, computer algorithms are utilized to answer **complex questions** (often via blackbox procedures).



- If we do not understand how A.I. makes decisions, we are not different from our ancestors.

Rational Decision Making



- Predictions should be **Precise** (no gibberish).
 - Need to study decision making under a math framework.
- Prediction should be **Data-driven**.
 - e.g. “sun rises up from west tomorrow” is not backed up by historical data.
- Takes **Cost** into consideration.
 - Cost of making a wrong decision may be different in tasks.
- Takes **Random nature** of Data into consideration!
 - Data generation/collection maybe noisy.

Statistical Decision Making

- We will see how **statistical** decision making exemplifies these guidelines.
- **Fun fact:** The way of taking randomness into account in decision making defines two distinct groups of statisticians: Frequentists and Bayesians.

Formal Notations

- x, y, z , scalars, $\mathbf{x}, \mathbf{y}, \mathbf{z}$, vectors.
- $\mathbf{x} \in R^d$, vector \mathbf{x} in d dimensional real-space.
- $x^{(i)}$, the i -th dimension of \mathbf{x} .
- X , a set
- $\mathbf{x}_i \in X$, the i -th member in X .
- $\mathbf{f}(\mathbf{x}) \in R^m$, function takes input vector \mathbf{x} and maps it into m dimensional real space.
- $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \in R^{b \times d}$, **matrices**, with b rows and d columns.
- “=” is equality, “:=” is definition.
 - $X := \{x_1, x_2\}$
 - $\sum_i \sum_j x_i y_j = \sum_j \sum_i x_i y_j$

Least Squares Regression

Regression Problem

- Regression is a common decision task.
- Predict outcome given some known inputs.
- For example,
 - Predict blood pressure given a patient's physical conditions.
 - Predict final year grade given a student's first-year scores.
 - etc.

Regression Problem

- **Input:** $\mathbf{x} \in R^d$
 - d -dimensional real-input,
 - e.g. weight, height, age, etc.
- **Output:** $y \in R$,
 - one dimensional real-output,
 - e.g. blood-pressure
- **The Problem:**
 - Given an input \mathbf{x} , predict its output.
- **Dataset** $D := \{(\mathbf{x}_i, y_i)\}_{i=1}^n$
 - Observed **pairs of inputs and outputs**.

Least Squares (LS)

$$\min_f \sum_{i \in D_0} [y_i - f(\mathbf{x}_i)]^2$$

- $f(\mathbf{x})$: **prediction function given \mathbf{x} .**
 - return a real-valued prediction
- $[\cdot]^2$: **square cost function.**
 - cost on difference between prediction and observed output
- $D_0 \subseteq D$: **training dataset.**
 - contains paired observations for tuning prediction f

Linear LS

$$\mathbf{w}_{\text{LS}} := \operatorname{argmin}_{\mathbf{w}} \sum_{i \in D_0} [y_i - f(\mathbf{x}_i; \mathbf{w})]^2$$

$$f(\mathbf{x}; \mathbf{w}) := \langle \mathbf{w}_1, \mathbf{x} \rangle + w_0, \mathbf{w} := \begin{bmatrix} \mathbf{w}_1 \\ w_0 \end{bmatrix}$$

- Solution: $\mathbf{w}_{\text{LS}} = (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{y}^\top$.
 - Suppose \mathbf{x} is a column vector.
 - $\mathbf{X} := \begin{bmatrix} \mathbf{x}_1, \dots, \mathbf{x}_n \\ 1, \dots, 1 \end{bmatrix} \in R^{(d+1) \times n}, \mathbf{y} = [y_1, \dots, y_n] \in R^n$.
 - **Proof: Homework**
- LS Prediction: $f(\mathbf{x}; \mathbf{w}_{\text{LS}})$.

Linear Least Squares (LS)

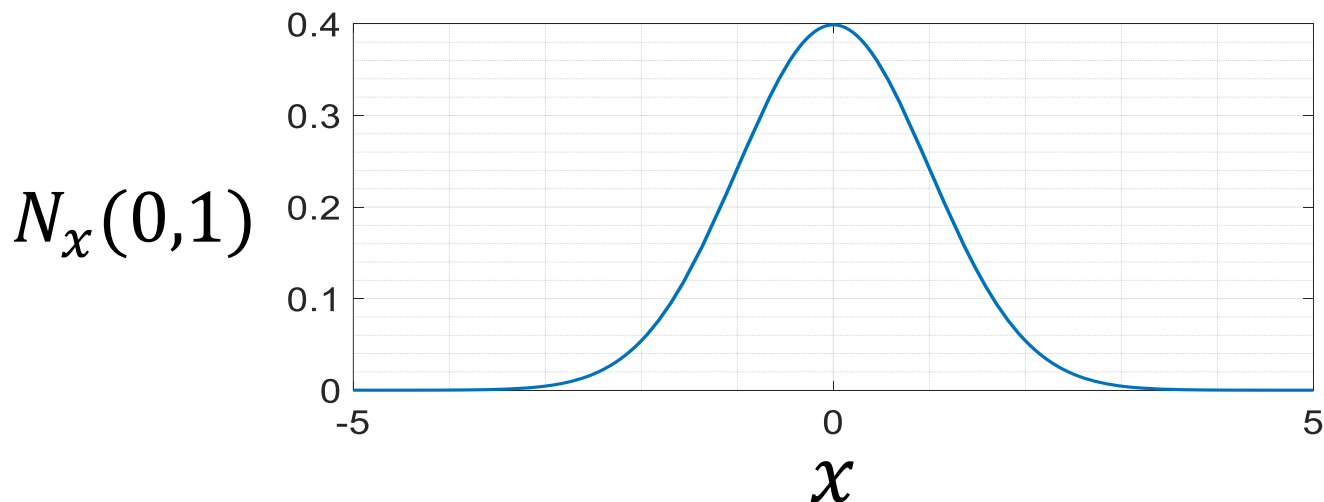
$$\mathbf{w}_{\text{LS}} := \operatorname{argmin}_{\mathbf{w}} \sum_{i \in D_0} [y_i - f(\mathbf{x}_i; \mathbf{w})]^2$$

$$f(\mathbf{x}; \mathbf{w}) := \langle \mathbf{w}_1, \mathbf{x} \rangle + w_0, \mathbf{w} := [\mathbf{w}_1, w_0]^\top$$

- LS is data-driven and uses squared function as its cost.
- **How does LS take randomness of dataset into account?**
- To answer this, we see LS from a probabilistic perspective.

Normal Distribution

- Random events of a Normal dist. happen on real domain.
- Normal dist. has a probability density function (PDF):
- $p(x|\mu, \sigma) := \frac{1}{Z(\sigma)} \exp \left[-\frac{(x-\mu)^2}{2\sigma^2} \right], Z(\sigma) = \sigma\sqrt{2\pi}, x \in R.$
- We use $N_x(\mu, \sigma^2)$ denote a Normal PDF. w.r.t. x .



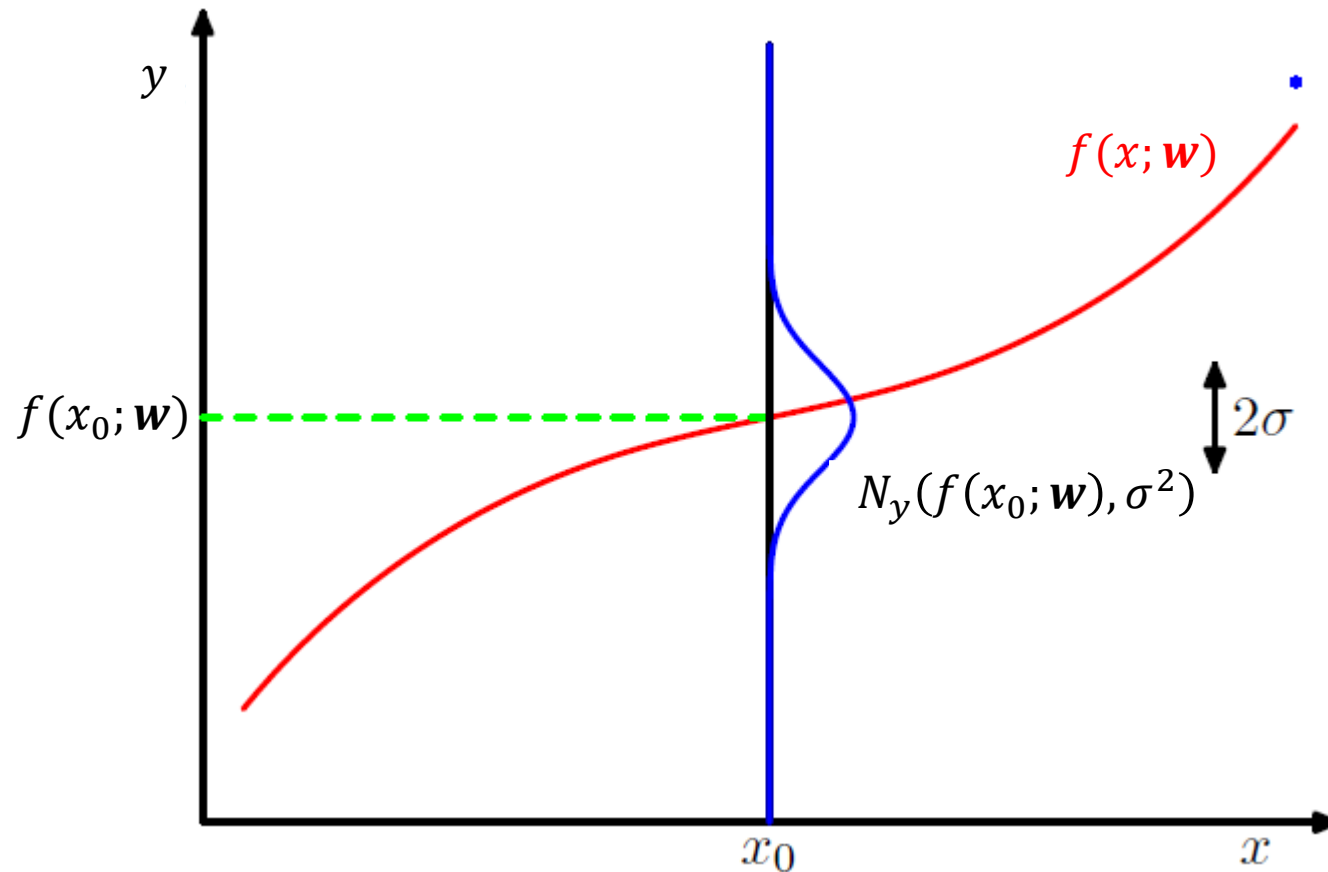
Probabilistic Modelling

PRML 1.2.5

- We express randomness of y using a prob. distribution.
- Given \mathbf{x} , we assume $p(y|\mathbf{x}, \mathbf{w}, \sigma) = N_y(f(\mathbf{x}; \mathbf{w}), \sigma^2)$.
 - y follows a Normal dist. with mean $f(\mathbf{x}; \mathbf{w})$ and var. σ^2 .
- This is only the model for a single y and \mathbf{x} pair.
 - We have a dataset of n (\mathbf{x}, y) pairs!
- By assuming (y_i, \mathbf{x}_i) are independent and identically distributed (IID), we can model
 - $p(y_1 \dots y_n | \mathbf{x}_1 \dots \mathbf{x}_n, \mathbf{w}, \sigma) = \prod_{i=1}^n N_{y_i}(f(\mathbf{x}_i; \mathbf{w}), \sigma^2)$.
 - Why?

LS from a probabilistic view

PRML Figure 1.16



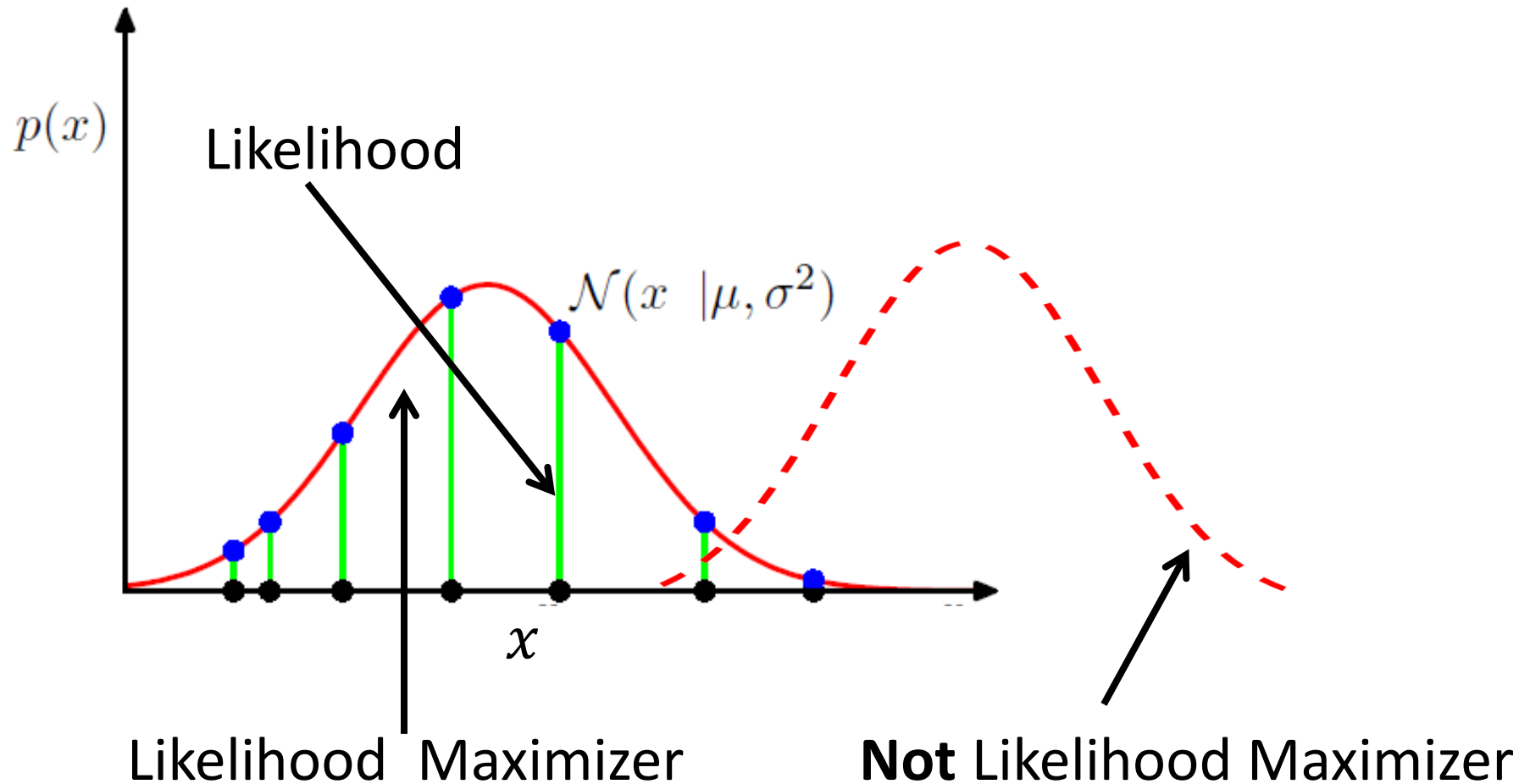
Q: How to determine \mathbf{w} and σ in a data-driven approach?

Maximum Likelihood Estimation (MLE)

PRML Figure 1.14

- PDF evaluated at observations are called **likelihood**.
- Given a dataset D , MLE maximizes (log) likelihood with respect to the unknown parameter θ .
- To determine parameter θ in $p(x|\theta)$:
- $\theta_{\text{ML}} := \underset{\theta}{\operatorname{argmax}} \log p(D|\theta) = \underset{\theta}{\operatorname{argmax}} \log p(x_1 \dots x_n|\theta)$
- Assuming $D := \{x_1 \dots x_n\}$ is IID
- $\theta_{\text{ML}} = \underset{\theta}{\operatorname{argmax}} \sum_i \log p(x_i|\theta)$

Likelihood and Maximizing Likelihood



LS from a probabilistic view

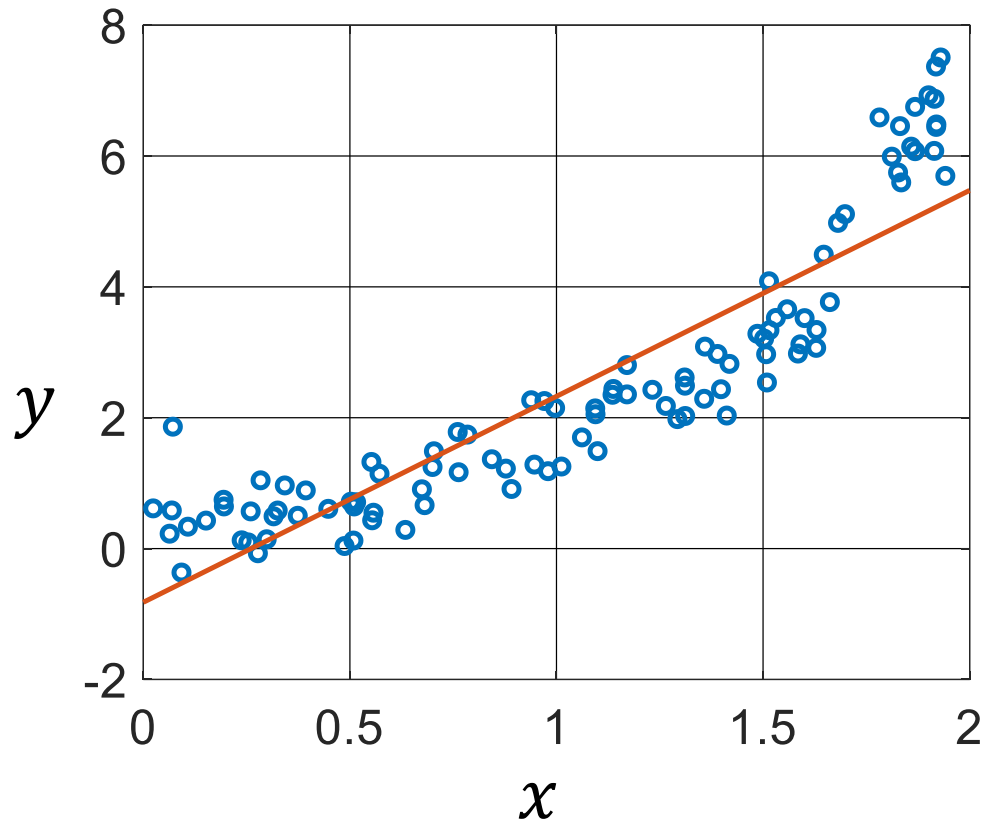
- We have
 - a probabilistic model of y given \mathbf{x} with unknown parameters
 - a dataset D_0
- We can perform MLE to find \mathbf{w}_{ML} !
- $\mathbf{w}_{\text{ML}} := \operatorname{argmax}_{\mathbf{w}} \log \prod_i^n N_{y_i}(f(\mathbf{x}_i; \mathbf{w}), \sigma^2)$
$$= \operatorname{argmax}_{\mathbf{w}} \left[\sum_{i=1}^n -\frac{(y_i - f(\mathbf{x}_i; \mathbf{w}))^2}{2\sigma^2} \right] - n \log \sigma \sqrt{2\pi}$$
$$= \operatorname{argmin}_{\mathbf{w}} \left[\sum_{i=1}^n (y_i - f(\mathbf{x}_i; \mathbf{w}))^2 \right]$$
- We can see $\mathbf{w}_{\text{ML}} = \mathbf{w}_{\text{LS}}$.

LS from a probabilistic view

- $\sigma_{\text{ML}} := \operatorname{argmax}_{\sigma > 0} \left[\sum_{i=1}^n -\frac{(y_i - f(x; \mathbf{w}))^2}{2\sigma^2} \right] - n \log \sigma \sqrt{2\pi}$
- $\sigma_{\text{ML}}^2 = \frac{1}{n} [y - f(\mathbf{x}; \mathbf{w}_{\text{ML}})]^2$
- This probabilistic view not only allows us to fit a prediction function f , but also the uncertainty of our prediction σ .
- This probabilistic view enables us to develop powerful regression tools on top of LS, which we will see in later.

LS with Feature Transform

- Linear LS only fits straight lines, which can be a problem if the relationship between y and x is non-linear.



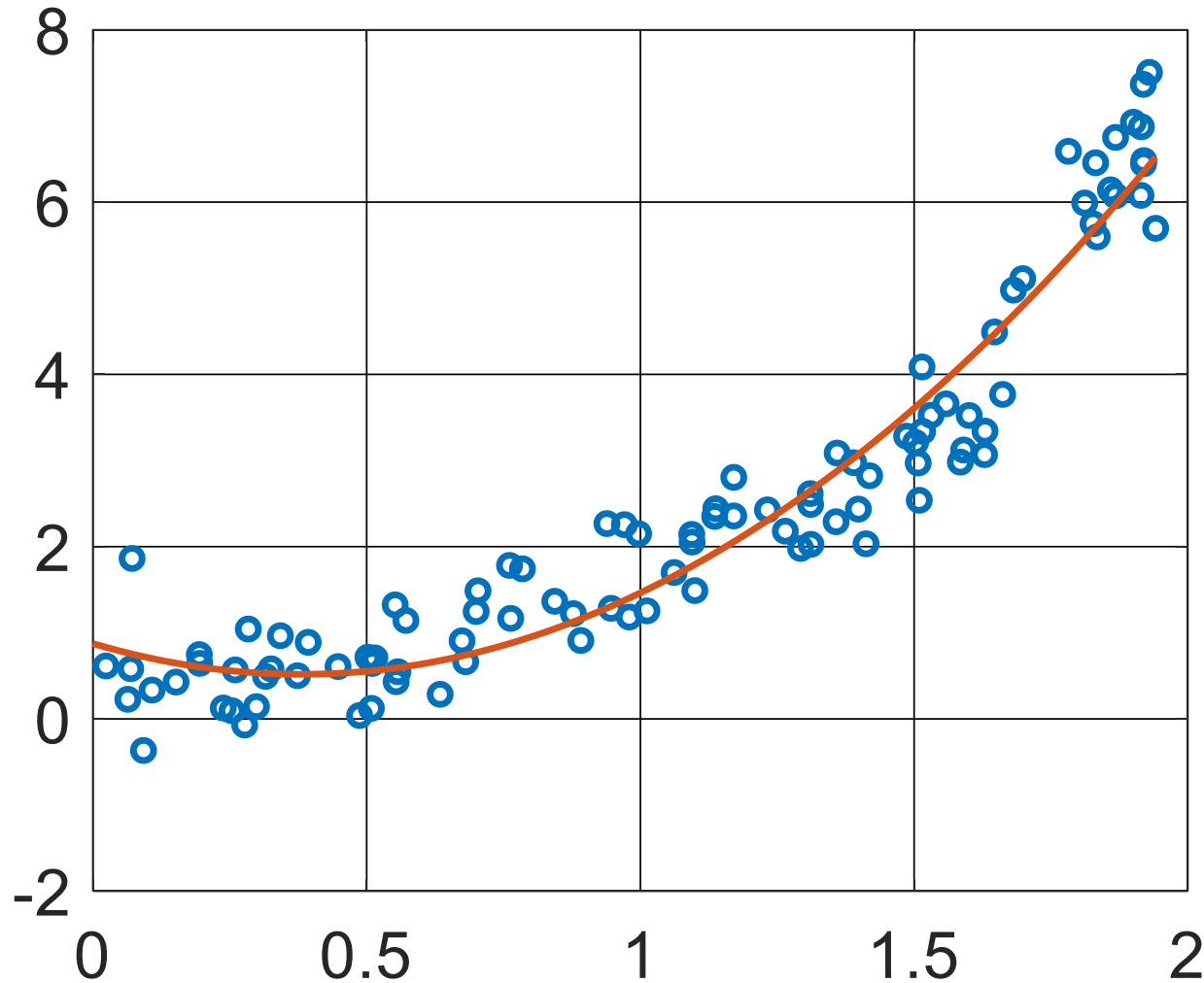
LS with Feature Transform

- It is easy to fit a nonlinear curve to our dataset, while maintaining the simple solution of linear LS.

$$\mathbf{w}_{\text{LS}} := \operatorname{argmin}_{\mathbf{w}} \sum_{i \in D_0} [y_i - f'(\mathbf{x}_i; \mathbf{w})]^2$$
$$f'(\mathbf{x}; \mathbf{w}) := \langle \mathbf{w}_1, \boldsymbol{\phi}(\mathbf{x}) \rangle + w_0, \mathbf{w} := \begin{bmatrix} \mathbf{w}_1 \\ w_0 \end{bmatrix}$$

- $\boldsymbol{\phi}(\mathbf{x}): R^d \rightarrow R^b$, is called a feature transform.
 - $\boldsymbol{\phi}(\mathbf{x}) := \mathbf{x}$, Linear transform.
 - $\boldsymbol{\phi}(\mathbf{x}) := [x, x^2, x^3, \dots, x^b]^\top$, Polynomial transform
- Solution: $\mathbf{w}_{\text{LS}} = (\boldsymbol{\phi}(\mathbf{X})\boldsymbol{\phi}(\mathbf{X})^\top)^{-1} \boldsymbol{\phi}(\mathbf{X})\mathbf{y}^\top$
- $\boldsymbol{\phi}(\mathbf{X}) := \begin{bmatrix} \boldsymbol{\phi}(\mathbf{x}_1), \dots, \boldsymbol{\phi}(\mathbf{x}_n) \\ 1, \dots, 1 \end{bmatrix} \in R^{(b+1) \times n}$,

LS with Polynomial Transform ($b = 2$)



LS with Feature Transform

$$\mathbf{w}_{\text{LS}} := \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i \in D_0} [y_i - f'(\mathbf{x}_i; \mathbf{w})]^2$$
$$f'(\mathbf{x}; \mathbf{w}) := \langle \mathbf{w}_1, \boldsymbol{\phi}(\mathbf{x}) \rangle + w_0, \mathbf{w} := [\mathbf{w}_1, w_0]^\top$$

- However, introducing complex feature transform in regression also opens cans of worms.
 - Overfitting
 - Curse of dimensionality
- Next lecture, we are going to see what are these problems and how to handle them using probabilistic methods.

Homework

- Prove $\mathbf{w}_{LS} = (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{y}^\top$
- The solution of \mathbf{w}_{LS} on page 15 is useless if $n < d$.
 - Why?
 - Can you find a solution to this problem?
- In what scenarios, the use of Normal distribution to model $p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma)$ on page 21 is a bad idea?
 - Find at least 2 scenarios and explain why.
- Prove $\mathbf{w}_{LS} = [\boldsymbol{\phi}(\mathbf{X})]^{-1} \mathbf{y}^\top$ if $\boldsymbol{\phi}(\mathbf{X})$ is symmetric and invertible.
- If we increase b of $\boldsymbol{\phi}(\mathbf{x})$ by 2-fold, by how many folds will the computation time of \mathbf{w}_{LS} increase?

Homework (Challenge)

- LS principle can be seen in many other machine learning problems outside of regression. Given a dataset $D := \{\mathbf{x}_i\}_{i=1}^n$, where $\mathbf{x} \in \mathbb{R}^d$. Consider the following objective:
- $\mathbf{w}' := \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^d, \langle \mathbf{w}, \mathbf{w} \rangle = 1} \sum_i \left| \mathbf{w} \mathbf{w}^\top \mathbf{x}_i - \mathbf{x}_i \right|^2$.
 - $\|\mathbf{a} - \mathbf{b}\|$: the Euclidean distance between two vectors \mathbf{a} and \mathbf{b} .
- Express \mathbf{w}' in terms of \mathbf{x}_i .
- What is the geometric interpretation of the obj. above?

Homework (Challenge)

- [Kullback-Leibler \(KL\) divergence](#) is a measure of dissimilarity between distributions and is expressed as:
- $KL[q, p] := \int q(x) \log \frac{q(x)}{p(x)} dx$
- If you have a probabilistic model $p(x|\theta)$ and you know your data is drawn from a probability distribution with density q , it makes sense to select your model parameter θ by $\min_{\theta} KL[q, p_{\theta}]$, **so that the fitted model is closest to the actual data distribution in terms of KL.**
- Q: What is the relationship between this model fitting objective and MLE? Under what assumptions, they are closely related?