

Machine Learning

INFQ612L, 440113450A

Spring Semester

Friday 17:00–18:40

IPS

WASEDA University

Prof. Shoji Makino



Machine Learning

Friday 17:00–18:40

1. 4/18

2. 4/25

3. 5/2

4. 5/9

5. 5/16

6. 5/23

7. 5/30

8. 6/6

9. 6/13

10. 6/20

11. 6/27

–. 7/4 No Lecture

–. 7/11 No Lecture

–. 7/18 No Lecture



-

At Zoom, set your name as:

Student ID, LAST_NAME, First_name

44251234, MAKINO, Shoji

At my class,

please turn on your camera

Machine Learning (2)(3)

Introduction to machine learning
and
Single regression

What is Machine Learning?

- Gives a computer the learning ability of humans
- Technology to obtain knowledge and predictions from data
 - Artificial intelligence (AI) ... Engineering pursuit of human intelligence
 - Statistics ... Theory of data analysis
 - Operations Research (OR) ... Theory of optimization and decision making
- Machine Learning
 - ≡ Almost synonymous with artificial intelligence in recent years

What Humans can do Easily

- Cut out the object of interest
- Generalize objects that are tied to past knowledge



Recognition/Decision

- Recognition: Replace data with concepts
 - Understand the object you see with your eyes as an old woman
- Decision: Make decisions based on data
 - Judge incoming email documents as spam/non-spam
- Human
 - Information processing in the brain (network of neurons)
 - The mechanism is not completely understood
- Machine Learning
 - Information processing of Recognition/Decision by computer

Example 1: Addressing Document Topic

- There are 10,000 data (documents, topics)
 - (Ghosn to resign as Renault director, International)
 - (Ichiro retires, a hole in the heart of an enemy player, sports)
 - (Japan-France relations "no particular influence", politics)
 - (Government's policy to deal with "Reiwa" reform, politics)
 - (Entertainer's rush divorce one after another, entertainment)
 - (Very thin Garigari Garixon, entertainment)
- New data comes
 - (Kei Nishikori keep the top 10, ?)
- What is the topic of this document?

Example 2: Grouping

- Which one is in the same group?

Google ロック

すべて 動画 画像 ニュース 地図 もっと見る 設定 ツール

コレクション セーフサーチ

サイズ 色 変更後の非営利目的での再使用が許可された画像 種類 時間 その他のツール リセット

iphone 壁紙 ライブ かわいい おすすめ おしゃれ スマホ ギター 猫 洋楽 イラスト バンド 背景 ディズニー グラム 携帯 おそろ

コンサート ロック 音楽 - Pixabayの無...
pixabay.com

ハードロック ハードロックカフェ ロック ...
pixabay.com

無料画像: 夜, 群集, スタジアム, ステージ, ...
pxhere.com

フリー写真画像: 群衆, ロック コンサート、...
pixnio.com

Mac OSX : IOS 7 ロック画面 風のスクリーン...
photozou.jp

iOS 7 : 音楽再生中のロッ...
photozou.jp

子供 再生 ロック - Pixabayの無料写真
pixabay.com

無料画像: 岩, ロックンロー...
pxhere.com

ピンク・フロイド, ロゴ, ロックンロール...
pexels.com

iOS 11 ロック画面 - 3 ; 画...
photozou.jp

フリー写真画像: ロック音楽、アーティスト...
pixnio.com

Machine Learning

	Supervised learning	Unsupervised learning
Predicted target: continuous value	Regression, Recommendation	Dimension reduction (PCA: Principal component analysis)
Predicted target: discrete value	Classification (Discrimination)	Clustering

Supervised Learning

- Supervised learning data
 - Known sample = (data, label)
 - Unknown sample = (data)
 - Predict the label (concept) of an unknown sample from known samples
- Data: Purchase history DB
 - Mr. A used X yen this month (continuous value)
- Data: Mail documents
 - Mail X is {spam, non-spam} (discrete value)
- Data: Images
 - Image X show "cat" (discrete value)

Unsupervised Learning

- Data on unsupervised learning
 - Samples = (Data)
- Data: News DB {Article 1, Article 2,...}
 - Articles 1, 3, and 5 are in a group (political news), and articles 2 and 4 are in a group (sports news) (discrete values)
 - Article 1 has a political component of 0.6 and an economic component of 0.4 (continuous values)
- Data: Server operation history DB
 - Traffic between 18:50 and 19:00 = abnormal, otherwise = normal (discrete value)
 - Traffic abnormal between 18:50 and 19:00 is 0.8, otherwise 0.01 (continuous value)

Flow of Machine Learning

Data



Toward the first budget compilation under the new era “Reiwa”, the national fiscal system council will consider measures to curb spending such as social security expenses as the budget scale continues to expand. We have set up a subcommittee to have intensive discussions.

The Financial System Council, which proposes to the Minister of Finance what the budget should be, began discussions on the 4th for the first budget formulation for the next fiscal year under “Reiwa”.

Features (vector representation of data)

$$X = \{x_1, x_2, \dots, x_N\},$$
$$x_i \in \mathbb{R}^D$$

Machine learning (supervised, unsupervised):
feature-to-concept mapping

Acquisition of concepts

Spam, Cat, Political News, etc.

Importance of Machine Learning

- Machine Learning
 - Capture the model behind it from a large amount of data and make predictions using its properties
- For the management of physical systems...
 - Identification and control of physical models is essential
- For the management of knowledge and information systems...
 - Information modeling and predictions based on it are effective
 - Machine learning is becoming an important elemental technology in knowledge and information systems, in today's world of large-scale contents

Regression Usage Scenarios: Wine Data Predicts Popularity

- Wine
- 10-dimensional features (see next page, real numbers), for 4,898 people
- Predict wine preferences (10 levels, integer numbers)
 - Predict continuous target value from continuous feature values
→ Regression problem



Data details

The physicochemical data statistics per wine type

Attribute (units)	Red wine			White wine		
	Min	Max	Mean	Min	Max	Mean
fixed acidity ($g(\text{tartaric acid})/dm^3$)	4.6	15.9	8.3	3.8	14.2	6.9
volatile acidity ($g(\text{acetic acid})/dm^3$)	0.1	1.6	0.5	0.1	1.1	0.3
citric acid (g/dm^3)	0.0	1.0	0.3	0.0	1.7	0.3
residual sugar (g/dm^3)	0.9	15.5	2.5	0.6	65.8	6.4
chlorides ($g(\text{sodium chloride})/dm^3$)	0.01	0.61	0.08	0.01	0.35	0.05
free sulfur dioxide (mg/dm^3)	1	72	14	2	289	35
total sulfur dioxide (mg/dm^3)	6	289	46	9	440	138
density (g/cm^3)	0.990	1.004	0.996	0.987	1.039	0.994
pH	2.7	4.0	3.3	2.7	3.8	3.1
sulphates ($g(\text{potassium sulphate})/dm^3$)	0.3	2.0	0.7	0.2	1.1	0.5
alcohol (% vol.)	8.4	14.9	10.4	8.0	14.2	10.4

Wine Data Popularity Forecast (Scenario)

Measurement



7.4;0.7;0;1.9;...;5
4.0;5 7.8;0.8; ...;8

...

Tasting



Learn predictive function
 $f(\text{sensor data}) = \text{popularity}$

Regression

$$t = w_0 + \sum_{i=1}^D w_i x_i$$

0.76;0.04;2.3;...;?

Measurement

Popularity
= 9

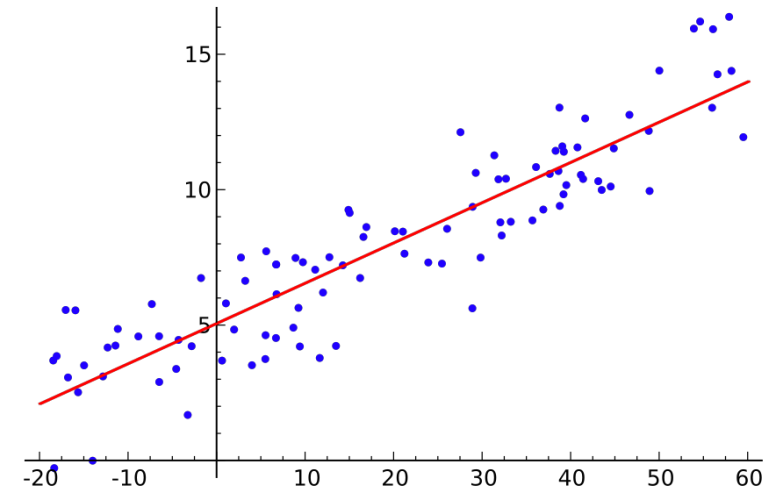
Predict popularity only
by measurement



New Wines

Single Regression

- Training data
 $(x_1, t_1), (x_2, t_2), \dots, (x_N, t_N)$
 - Feature value: x_i Observable data
 - Target value: t_i Prediction target

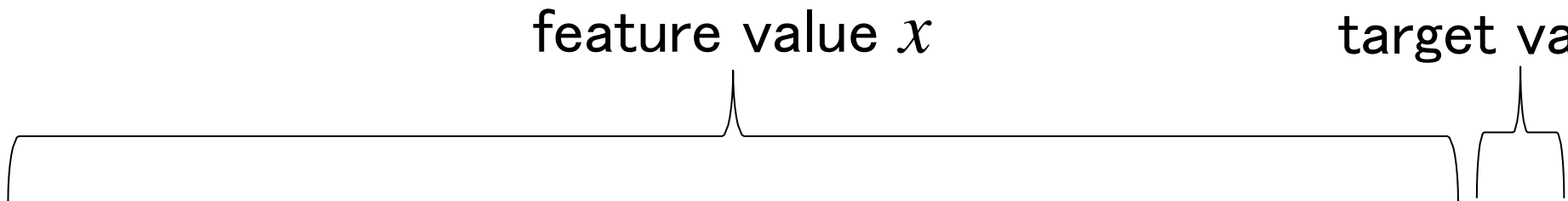


- Purpose
 - Find a linear model between features and target values from a large number of training data (learning)
 - Have a linear model and predict the target value, when a feature with an unknown target value is given (prediction)

Linear Model

- Linear relationship between target value t and feature value x
 - Example: predict the lifespan t from weight x $t = wx + b$
 - Example: predict the total amount of assets t from age x
- Parameters w
 - Strength of the relationship between feature value x and target value t (the higher the absolute value, the stronger the relationship)
 - Positive sign: The larger the feature value, the larger the target value
 - Example: As you get older, your total wealth increases
 - Negative sign: The smaller the feature value, the larger the target value
 - Example: If you gain weight, your lifespan will decrease
- Bias b
 - Target value t when feature value x is 0

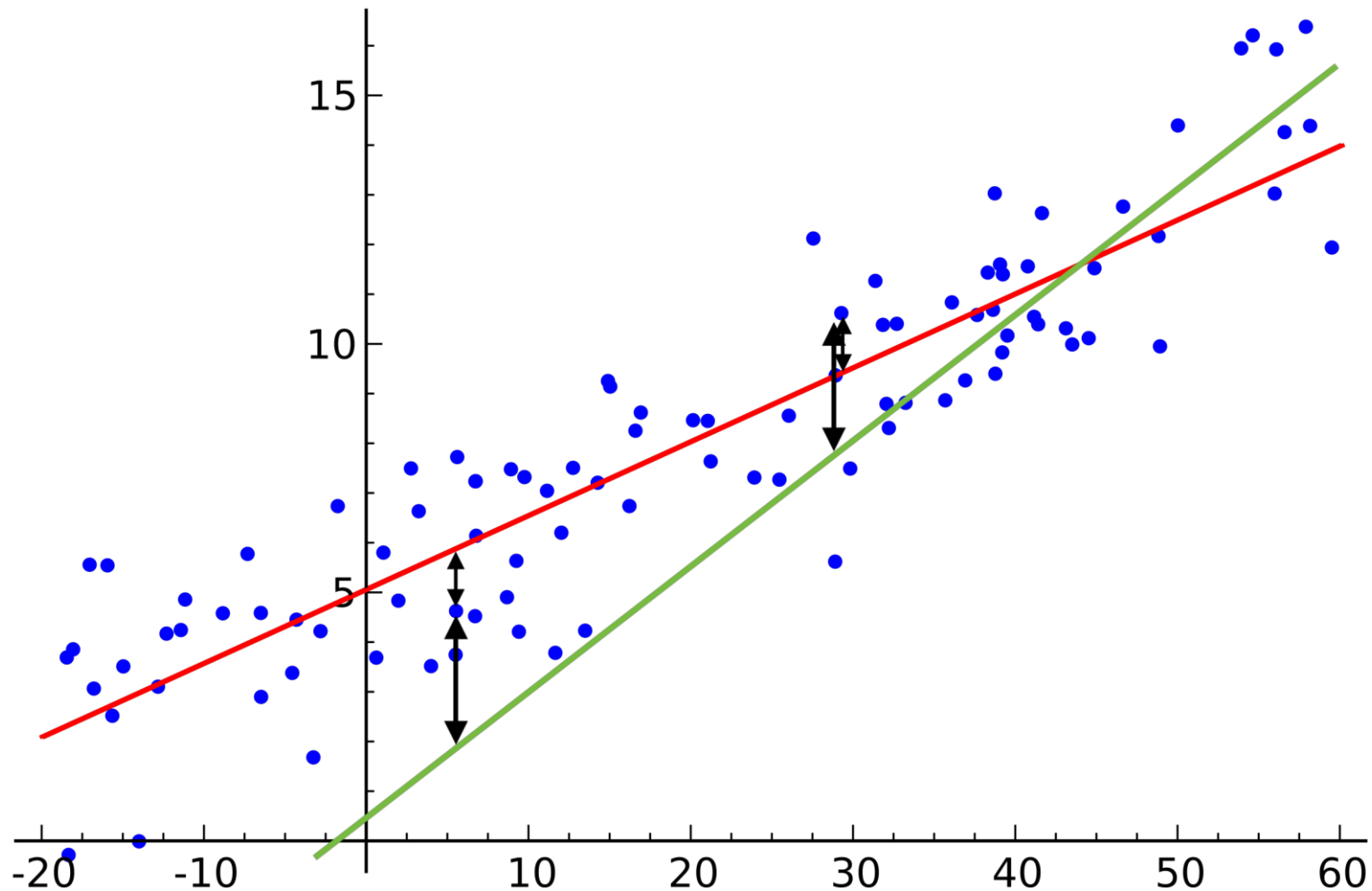
Wine Data



	feature value x											target value t
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

- Consider a linear equation that predicts the target value t from the feature value x
 - $\text{quality} = w_1 * \text{fixed_acidity} + b_1$
 - $\text{quality} = w_2 * \text{volatile_acidity} + b_2$

What are the parameters that give good predictions?



Mean Square Error

- Error: Difference between the actual target value and the predicted target value
error = $|t_i - (wx + b)|$

- Mean square error in the model $wx + b$
 - Average squared error for all training data

$$E(w) = \frac{1}{N} \sum_{i=1}^N (t_i - (wx_i + b))^2$$

- Good model = model that minimizes mean square error
- How to find the best model?
- Find w, b that minimizes the following

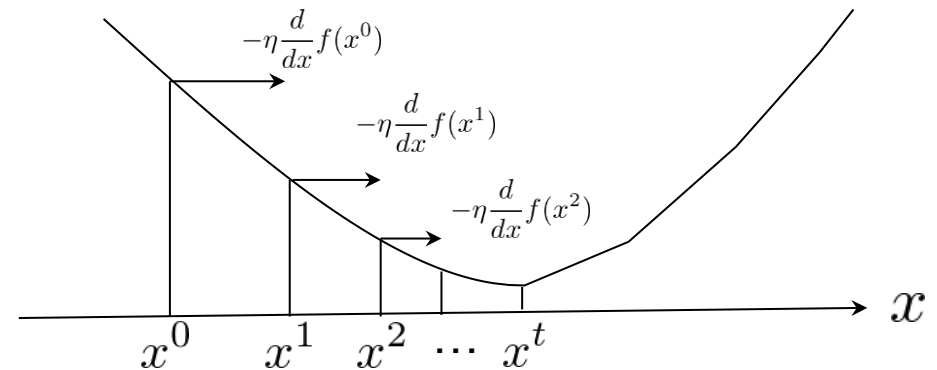
$$\min_{w,b} \frac{1}{N} \sum_{i=1}^N (t_i - (wx_i + b))^2$$

...How?

Gradient Descent Method

- Problem: find x that minimizes $f(x)$ $\min_x f(x)$
 - Function f is differentiable
- Algorithm
 1. $t = 0, x^0$ randomly initialize
 2. $x^{t+1} \leftarrow x^t - \eta \frac{d}{dx} f(x^t)$
 3. $t \leftarrow t + 1$
 4. $|x^{t+1} - x^t| < \epsilon$

Step size parameter η : How much to update in one step
Convergence judgment parameter ϵ



Exercise 1 :

Minimize the mean square error of single regression by the gradient descent method

$$E(w) = \frac{1}{N} \sum_{i=1}^N (t_i - (wx_i + b))^2$$

1. Find the derivative of the mean squared error $E(w)$ with parameters w and b .
2. Find the update equation for Step 2 of the Gradient Descent Method. Let the initial parameters be (w_0, b_0) , the t -th update parameters be (w_t, b_t) , and the step size parameter be η .

Machine Learning (2)(3)

Introduction to machine learning
and
Single regression

Target Value and Feature Value

- Target value: Value you want to predict
 - It is often decided by itself according to the purpose
 - Often a value that is expensive to acquire
- Feature: value that is (possibly) related to the prediction of the target value
 - It is usually not clear what data is a good feature
 - Good features lead to good prediction accuracy
 - We want to choose a value that does not cost to acquire
 - We want to obtain a target value that is costly to acquire at low cost using regression
 - Finding excellent and inexpensive features is very important

Data Type

- Numerical attribute
 - Attributes represented by scalar/vector
 - Example: age, income
- Order attribute
 - Attribute with ordinal relationship
 - Example: grade evaluation (A, B, C, D)
- Categorical attributes
 - Non-ordered attributes
 - Example: blood type

Data Conversion

- Machine learning assumes that features are represented by vectors
- Convert order attributes and categorical attributes to numerical attributes
- Order attribute:
 - Convert order-saved scalar value
 - Example: $A^+ \rightarrow 5$, $A \rightarrow 4$, ..., $D \rightarrow 1$
- Categorical attributes
 - 1-of-k conversion (one-hot encoding)
 - Examples: Type A (0,0), Type B (0,1), Type O (1,0), Type AB (1,1)
 - If there are N types of attribute values, $k = \log_2 N$

What to look for in the data?

- Convert a person's place of residence (47 prefectures) into a feature
- Predict the price of that person's house
 - Land prices should be higher in prefectures with higher population densities
 - Convert residential prefecture to population density (numerical attribute)
 - Tokyo \rightarrow 6,309, Osaka \rightarrow 4,631, ..., Hokkaido 67
- Predict how many mufflers a person has
 - Should be more to live in the north
 - Convert the prefecture you live into an ordinal attribute about latitude
 - Hokkaido \rightarrow 1, Aomori \rightarrow 2, ..., Okinawa \rightarrow 47
- Predict the person's Sweets consumption
 - Maybe there are prefectures that consume a lot of Sweets
 - 1-of-k encoding
 - Hokkaido = (1,0,0,..., 0), Aomori = (0,1,..., 0,0,0), ..., Okinawa = (0,0,..., 0,0,1)

Adult Dataset

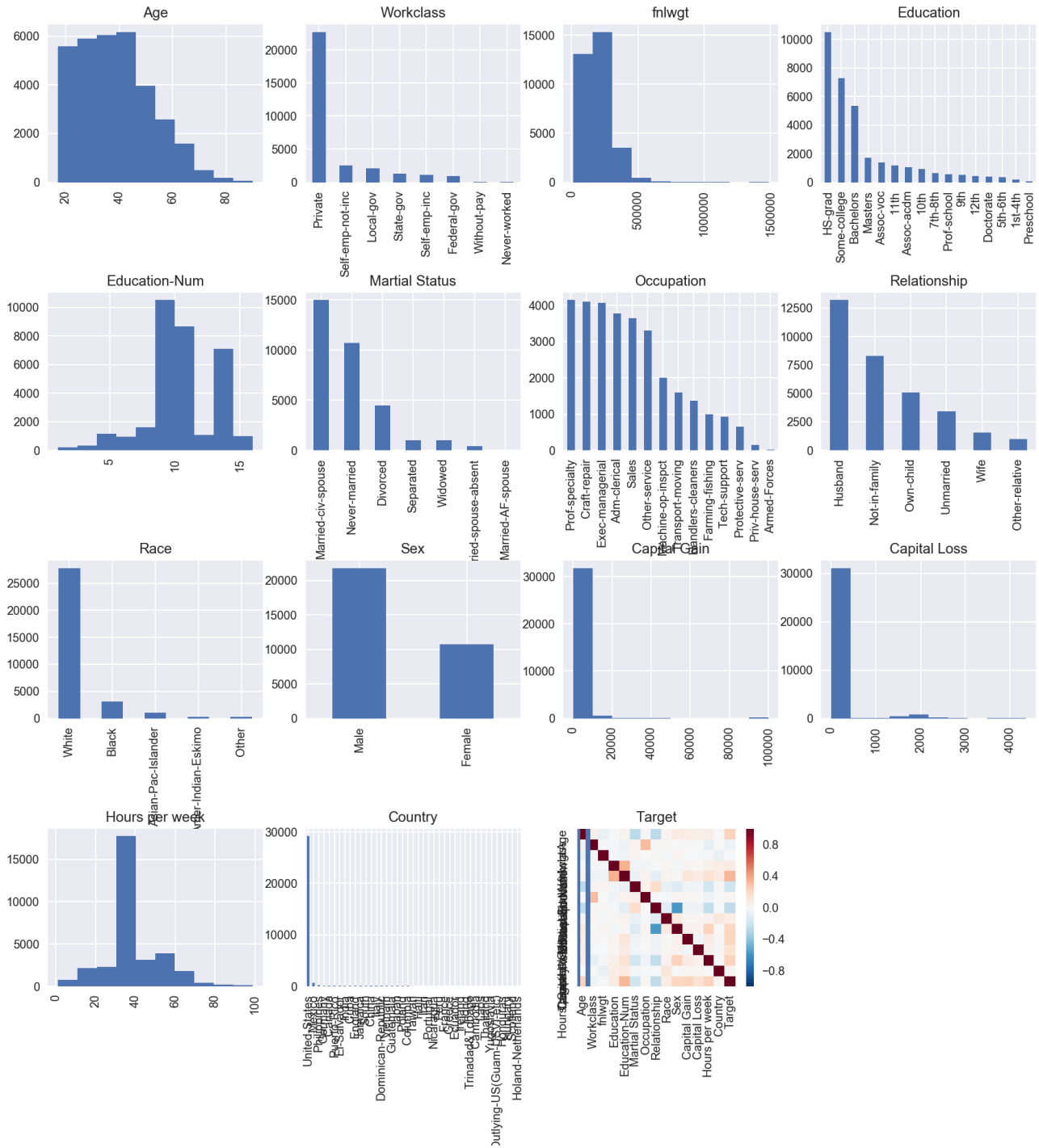
- Test data frequently used to evaluate machine learning performance
- Various attribute data of 50,000 individuals
- Predict whether the individual's income will exceed \$50,000

Adult Dataset Data Example

	Age	Workclass	fnlwgt	Education	Education-Num	Marital Status	Occupation	Relationship	Race	Sex	Capital Gain	Capital Loss	Hours per week	Country	Target
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
5	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
6	49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
7	52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
8	31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
9	42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K

Adult Dataset

- age: continuous.
- workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- fnlwgt: continuous.
- education: Doctorate, Prof-school, Masters, Bachelors, Assoc-acdm, Assoc-voc, Some-college, HS- grad, 12th, 11th, 10th, 9th, 7th-8th, 5th-6th, , 1st-4th, Preschool.
- education-num: continuous.
- marital-status: Married-civ-spouse, Married-AF-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent
- occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv- house-serv, Protective-serv, Armed-Forces.
- relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- sex: Female, Male.
- capital-gain: continuous.
- capital-loss: continuous.
- hours-per-week: continuous.
- native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying- US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El- Salvador, Trinidad&Tobago, Peru, Hong, Holand-Netherlands.



Preprocessing Data

- Dealing with missing values
 - Remove samples containing missing values
 - Missing values are complemented by the mean/median value of the feature
 - why?
- Correspondence to outliers
 - Take a bird's eye view and remove samples with extreme outliers (measurement errors, typos, possible exceptions)
 - why?
- Scaling
 - Scaling each feature to mean 0 and standard deviation 1
 - Scale so that the minimum value is 0 and the maximum value is 1
 - why?

Notation and Formula

- Notation is important:

- Scalar

$$x \in \mathbb{R}$$

- d -dimensional vector

$$\mathbf{x} \in \mathbb{R}^d$$

- n row m column matrix

$$\mathbf{X} \in \mathbb{R}^{n \times m}$$

- Transpose vector

$$\mathbf{x}^T$$

Frequently Used Notation

- Inner product $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^T \mathbf{b} = \mathbf{b}^T \mathbf{a}$
- Matrix vector product

$$\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_D, \mathbf{a} \in \mathbb{R}^D$$
$$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_D)$$

Then,

$$\mathbf{X}^T \mathbf{a} = \begin{pmatrix} \mathbf{x}_1^T \mathbf{a} \\ \mathbf{x}_2^T \mathbf{a} \\ \vdots \\ \mathbf{x}_D^T \mathbf{a} \end{pmatrix}$$

Differentiation of Multivariable Functions

- Function $f : \mathbb{R}^D \rightarrow \mathbb{R}$ $f(\mathbf{x}) = a$
- Derivative at a variable (x_d) $\frac{\partial f}{\partial x_d}$
- Derivatives (gradients) for all variables

$$\frac{\partial f}{\partial \mathbf{x}} = \nabla_{\mathbf{x}} f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{pmatrix}$$

Chain Rule

- Differential chain rule
- Chain rule holds even for multivariable functions

$$\frac{df(g(x))}{dx} = \frac{df(g(x))}{dg(x)} \frac{dg(x)}{dx}$$

$$\frac{\partial f(g(\mathbf{x}))}{\partial \mathbf{x}} = \frac{\partial f(g(\mathbf{x}))}{\partial g(\mathbf{x})} \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}}$$

- Example

$$f(\mathbf{x}) = (\mathbf{a}^T \mathbf{x} + b)^2$$

$$\frac{\partial f}{\partial \mathbf{x}} = 2(\mathbf{a}^T \mathbf{x} + b) \frac{\partial}{\partial \mathbf{x}} (\mathbf{a}^T \mathbf{x} + b)$$

$$= 2(\mathbf{a}^T \mathbf{x} + b) \mathbf{a}$$

Chain Rule of Multivariable Functions

- Multivariable function $f(\mathbf{g}) = f(g_1, g_2, \dots, g_M)$
- Each g_k is a function of x $g_k(x), k \in \{1, \dots, M\}$
- then

$$\frac{\partial f}{\partial x} = \sum_{k=1}^M \frac{\partial f}{\partial g_k} \frac{\partial g_k}{\partial x}$$

Exercise 2:

Matrix and gradient

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$$

とする。

1. $\mathbf{x}^T \mathbf{y} = \mathbf{y}^T \mathbf{x} = x_1 y_1 + \dots + x_n y_n$ を示せ。
2. $\mathbf{x}^T \mathbf{A} \mathbf{x}$ を成分 $(x_i, a_{i,j})$ で表せ。
3. $\frac{\partial}{\partial \mathbf{x}} \mathbf{a}^T \mathbf{x} = \mathbf{a}$ を示せ。
4. $\frac{\partial}{\partial \mathbf{x}} (\mathbf{a}^T \mathbf{x} + b)^2$ を求めよ。

Machine Learning (2)(3)

Introduction to machine learning
and
Single regression

Machine Learning

INFQ612L, 440113450A

Spring Semester

Friday 17:00–18:40

IPS

WASEDA University

Prof. Shoji Makino

