L4: Logistic Regression

Shan Wang

Lingnan College, Sun Yat-sen University

2020 Data Mining and Machine Learning LN3119

https://wangshan731.github.io/DM-ML/

Last lecture

- Linear regression method
  - Model: $y = \boldsymbol{\theta}'\boldsymbol{x}$
  - Strategy
    - Least squared error: $\min \frac{1}{N}\sum_{i=1}^{N}(y_i - \boldsymbol{\theta}'\boldsymbol{x}_i)^2$
    - Maximum likelihood: $\max \sum_{i=1}^{N} \log P(y_i|\boldsymbol{x}_i, \theta)$
  - Algorithm
    - Normal equation: $\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}'\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}'\boldsymbol{y}$
    - Gradient descent method: $\boldsymbol{\theta}_{\text{new}} = \boldsymbol{\theta}_{\text{old}} - \eta \frac{\partial R(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$
- Regularization
  - Lasso: $\lambda||\boldsymbol{\theta}||_1$
  - Ridge: $\lambda||\boldsymbol{\theta}||_2^2$
    - Relationship between Ridge and MAP
- Application: quality of wine

Regression vs Classification

- **Linear regression** would predict a **continuous** outcome
  - $y = \boldsymbol{\theta}'\boldsymbol{x}$
- To predict the quality of care
  - The dependent variable is modelled as a binary variable
  - 1 if low-quality care, 0 if high-quality care
- This is a **categorical variable**
  - Typically a small number of possible outcomes, 2 (low-quality care and high-quality care) in this case
- How can we extend the idea of linear regression to situations where the outcome variable is categorical?
  - Only want to predict 1 or 0
  - Could round outcome to 1 or 0
  - But we can do better with **logistic regression**

Course outline

- Supervised learning
  - Linear regression
  - Logistic regression
  - SVM and kernel
  - Tree models
- Unsupervised learning
  - Clustering
  - PCA
  - EM
- Deep learning
  - Neural networks
  - Convolutional NN
  - Recurrent NN
- Reinforcement learning
  - MDP
  - ADP
  - Deep Q-Network

This Lecture

- Classification problem
- Logistic regression method
  - Model
  - Strategy
  - Algorithm
- Prediction & evaluation
- Multi-class logistic regression
- Application: diabetes care

Reference: VE 445, Shuai LI (SJTU); OPIM 326 Daniel Zheng (SMU)

Classification problem

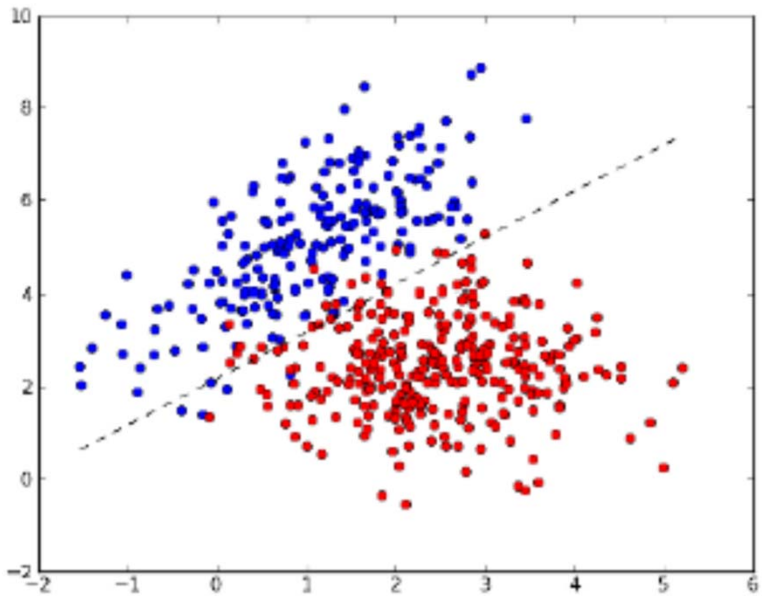
What is classification?

- Compared to regression problem, which predicts the labels from many numerical features
- Many applications
  - Spam Detection: Predicting if an email is Spam or not based on word frequencies
  - Credit Card Fraud: Predicting if a given credit card transaction is fraud or not based on their previous usage
  - Health: Predicting if a given mass of tissue is benign or malignant
  - Marketing: Predicting if a given user will buy an insurance product or not
  - …

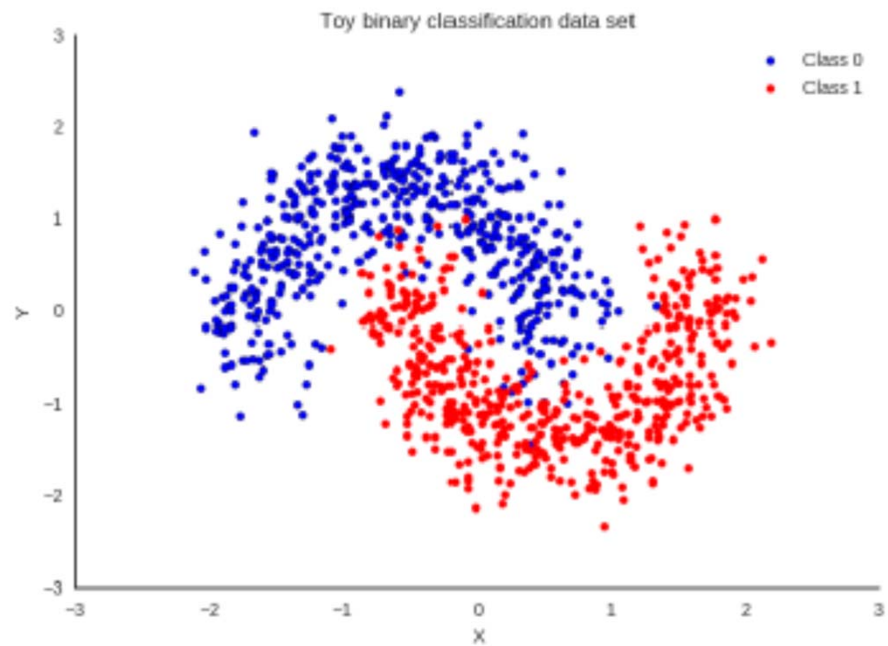
Classification problem definition

- Given:
  - A description of an instance $x \in X$
  - A fixed set of categories: $C = \{c_1, c_2, \dots, c_K\}$
- Determine:
  - The category of x: $f(x) \in C$ where $f(x)$ is a categorization function whose domain is X and whose range is C
  - If the category set binary, i.e. $C = \{0, 1\}$ ({false, true}, {negative, positive}) then it is called binary classification

Binary classification



Linearly separable



Nonlinearly separable

Linear discriminative model

- Modeling the dependence of unobserved variables on observed ones
- a.k.a. conditional models
- Deterministic: $y = f_\theta(x)$
- Probabilistic: $p_\theta(y|x)$

- For binary classification
  - $p_\theta(y = 1|x)$
  - $p_\theta(y = 0|x) = 1 - p_\theta(y = 1|x)$

Logistic regression model

What is logistic regression

- Logistic regression is a binary classification model
  - $p_\theta(y=1|\boldsymbol{x}) = \sigma(\boldsymbol{\theta}'\boldsymbol{x}) = \frac{e^{\boldsymbol{\theta}'\boldsymbol{x}}}{1+e^{\boldsymbol{\theta}'\boldsymbol{x}}}$
  - $p_\theta(y=0|\boldsymbol{x}) = \frac{1}{1+e^{\boldsymbol{\theta}'\boldsymbol{x}}}$
  - $\sigma(x)$ is the logistic function or the sigmoid function

- Interpretation
  - Odds: $\frac{p}{1-p}$
  - $p = \frac{e^{\boldsymbol{\theta}'\boldsymbol{x}}}{1+e^{\boldsymbol{\theta}'\boldsymbol{x}}} \leftrightarrow \log(\frac{p}{1-p}) = \boldsymbol{\theta}'\boldsymbol{x}$
    - Logistic regression is a linear regression for the log odds

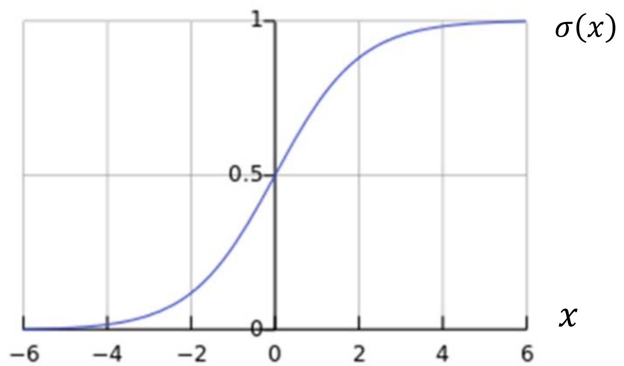
*Intercept term is omitted in this lecture, you may imagine the first term in $\boldsymbol{x}$ is always 1.

Logistic function properties

- Properties for the logistic function
  - $\sigma(x) = \frac{e^x}{1+e^x}$
    - Bounded in $(0,1)$
    - $\sigma(x) \to 1$ when $x \to \infty$
    - $\sigma(x) \to 0$ when $x \to -\infty$
  - $\sigma'(x) = \frac{\partial \frac{e^x}{1+e^x}}{\partial x}$

$$= \frac{e^x(1+e^x) - e^x e^x}{(1+e^x)^2} = \frac{e^x}{1+e^x} \times \frac{1}{1+e^x}$$

$$= \sigma(x)(1 - \sigma(x))$$



Logistic regression strategy

- Goal: we want to choose the right $\boldsymbol{\theta}$, to make the best prediction
- Which loss function to use?

Entropy

- Entropy
  - A measure of the uncertainty
  - Suppose the are K classes, class k with p_k
    - Entropy$= -\sum_{k=1}^{K} p_k \log p_k$
  - Uniform distribution has maximum entropy

- Cross entropy
  - To calculate the difference between two probability distributions
    - E.g., The true distribution and predicted distribution
  - Cross entropy$= -\sum_{k=1}^{K} p_k \log q_k$
    - p_k: true label distribution
    - q_k: predicted label distribution

Cross entropy loss for logistic regression

- Loss function for data point $(\boldsymbol{x}, y)$ with prediction model $p_\theta(\cdot\,|x)$ is

$$L(y, \boldsymbol{x}, p_\theta) =$$
$$= -1_{y=1} \log p_\theta(1|\boldsymbol{x}) - 1_{y=0} \log p_\theta(0|\boldsymbol{x})$$
$$= -y \log p_\theta(1|\boldsymbol{x}) - (1-y) \log(1 - p_\theta(1|\boldsymbol{x}))$$

- Where $p_\theta(1|\boldsymbol{x}) = \sigma(\boldsymbol{\theta}'\boldsymbol{x})$

- Minimize the empirical error $\hat{R}(\boldsymbol{\theta})$

$$\hat{\boldsymbol{\theta}} = \operatorname{argmin} -\frac{1}{N}\sum_{i=1}^{N}[y_i \log p_\theta(1|\boldsymbol{x}_i) + (1-y_i)\log(1 - p_\theta(1|\boldsymbol{x}_i))]$$

Logistic regression algorithm

Solve the $\hat{\boldsymbol{\theta}}$

- $\hat{R}(\boldsymbol{\theta})$ is convex in $\boldsymbol{\theta}$
- Gradient descent:
- $\frac{\partial L(y,x,p_{\theta})}{\partial \boldsymbol{\theta}} = -y\frac{1}{\sigma(\boldsymbol{\theta}'\boldsymbol{x})}\sigma(\boldsymbol{\theta}'\boldsymbol{x})\big(\mathbf{1}-\sigma(\boldsymbol{\theta}'\boldsymbol{x})\big)\boldsymbol{x}$

$$-(1-y)\frac{-1}{1-\sigma(\boldsymbol{\theta}'\boldsymbol{x})}\sigma(\boldsymbol{\theta}'\boldsymbol{x})\big(\mathbf{1}-\sigma(\boldsymbol{\theta}'\boldsymbol{x})\big)$$

$$= -\big(\boldsymbol{y}-\sigma(\boldsymbol{\theta}'\boldsymbol{x})\big)\boldsymbol{x}$$

$\sigma'(x) = \sigma(x)(1-\sigma(x))$

- $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \boldsymbol{\eta}\big(\boldsymbol{y}-\sigma(\boldsymbol{\theta}'\boldsymbol{x})\big)\boldsymbol{x}$

Prediction & evaluation

Label prediction

- Logistic regression provides the probability
  - $p_\theta(y = 1|\boldsymbol{x}) = \sigma(\boldsymbol{\theta}'\boldsymbol{x}) = \frac{e^{\boldsymbol{\theta}'\boldsymbol{x}}}{1+e^{\boldsymbol{\theta}'\boldsymbol{x}}}$
  - $p_\theta(y = 0|\boldsymbol{x}) = \frac{1}{1+e^{\boldsymbol{\theta}'\boldsymbol{x}}}$
- The final label of an instance is decided by setting a threshold h
  - If $p_\theta(y = 1|\boldsymbol{x}) > h$, predict 1
  - Otherwise, predict 0

How to choose threshold?

- In general, 0.5 is a good choose

- If you have precision-recall trade-off
  - Higher threshold
    - More False Negative
    - Less False Positive
    - Higher precision
    - Lower recall
  - Lower threshold
    - Less False Negative
    - More False Positive
    - Lower precision
    - Higher recall

		Prediction	
		1	0
Label	1	True Positive	False Negative
	0	False Positive	True Negative

- **Precision**: the ratio of true class 1 cases in those with prediction 1

$$\text{Prec} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

		Prediction	
		1	0
Label	1	True Positive	False Negative
	0	False Positive	True Negative

- **Recall**: the ratio of cases with prediction 1 in all true class 1 cases

$$\text{Rec} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Model evaluation

- Accuracy
  - How many are correctly classified
- Recall
  - How many positive ones are correctly classified
- Precision
  - How many are correctly classified among those are labelled as positive
- F1 measure

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

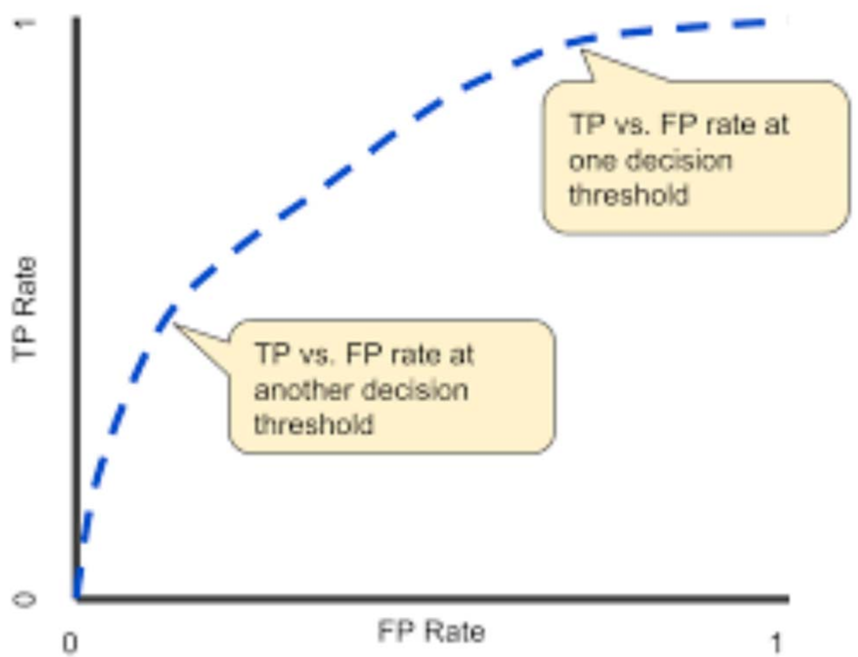
$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - measure = \frac{2TP}{2TP + FP + FN}$$

ROC curve

- ROC curve:
  - Receiver Operating Characteristic (ROC) Curve
  - X axis: false positive rate (1-specificity)=FP/(TN+FP)
  - Y axis: true positive rate (recall/sensitivity)=TP/(TP+FN)
  - A performance measurement for classification problem at various thresholds settings



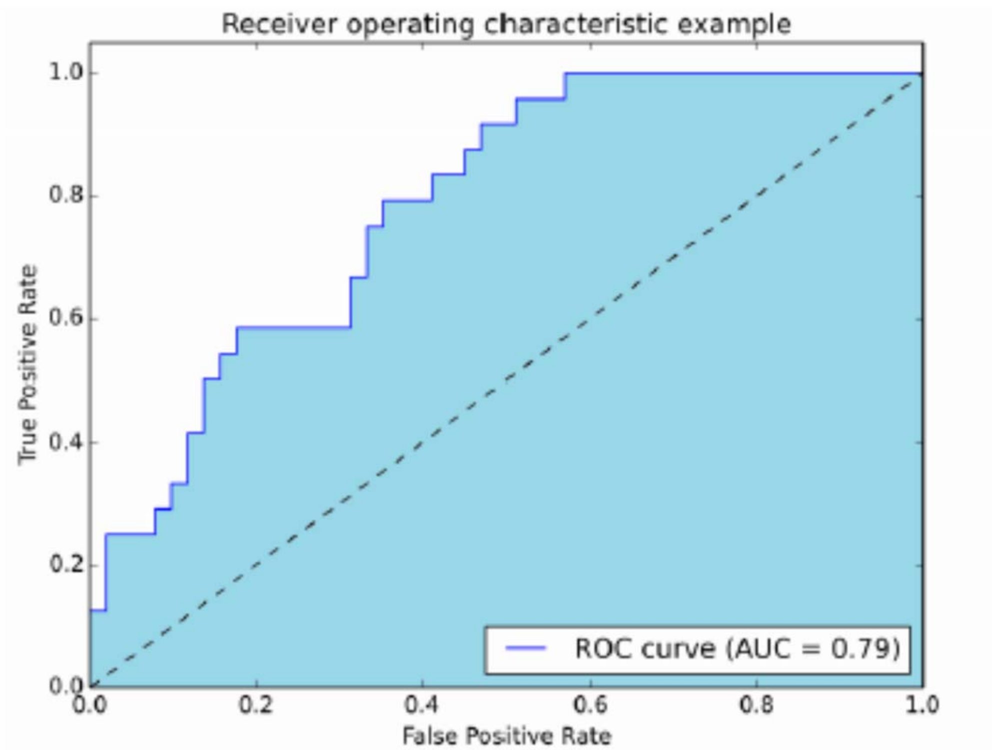
Bottom left corner, threshold=1, every thing is predicted to be negative, so TPR=0 and FPR=0

Decrease the threshold, more true positives, more false positives

Top right corner, threshold=0, every thing is predicted to be positive, so TPR=1 and FPR=1

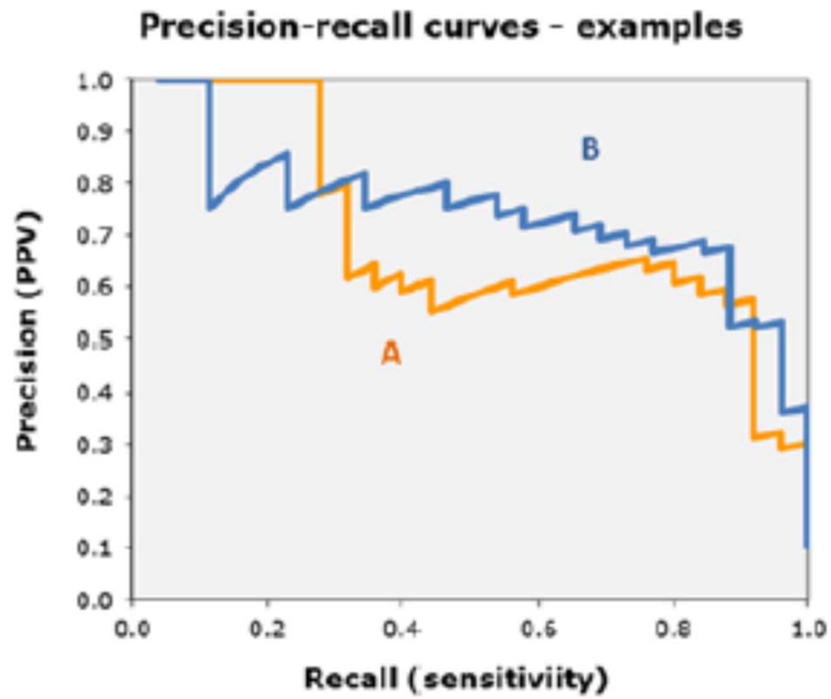
AUC

- Area Under ROC Curve (AUC)
  - The higher, the better
  - Tells how much the model is capable of distinguishing between classes
  - Perfect classifier get AUC=1 and random classifier get AUC = 0.5

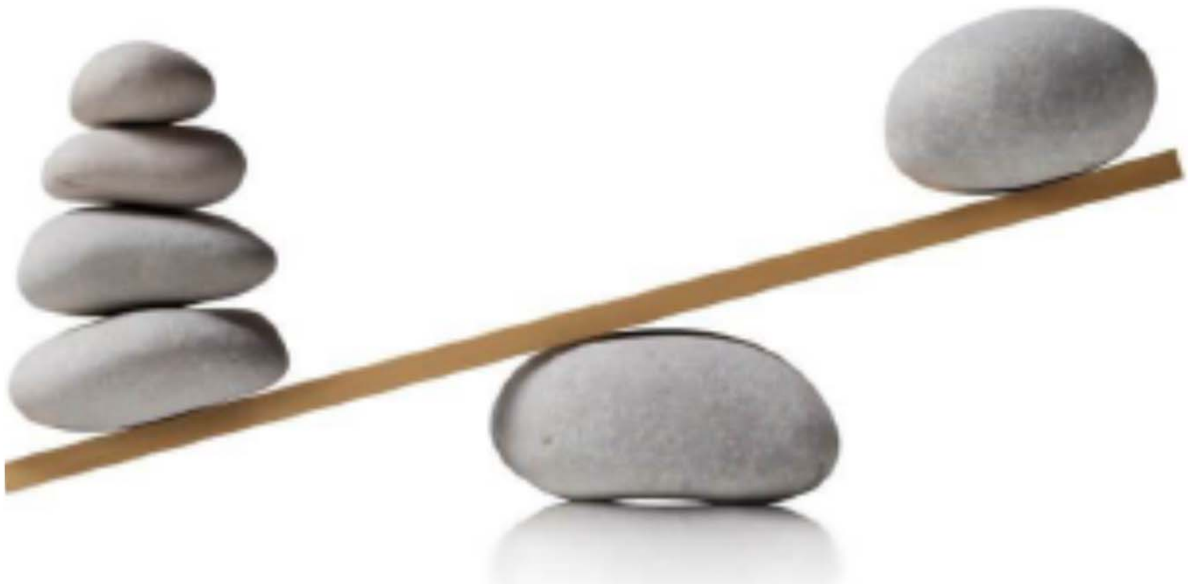


PR curve and AUPR

- PR curve:
  - The precision recall curve
  - X axis: recall=TP/(TP+FN)
  - Y axis: precision=TP/(TP+FP)



- AUPR
  - Area Under PR curve
  - Can handle imbalanced datas
  - another plot to measure the performance of binary classifier
  - Usually, the classifiers gets lower AUPR value than AUC value



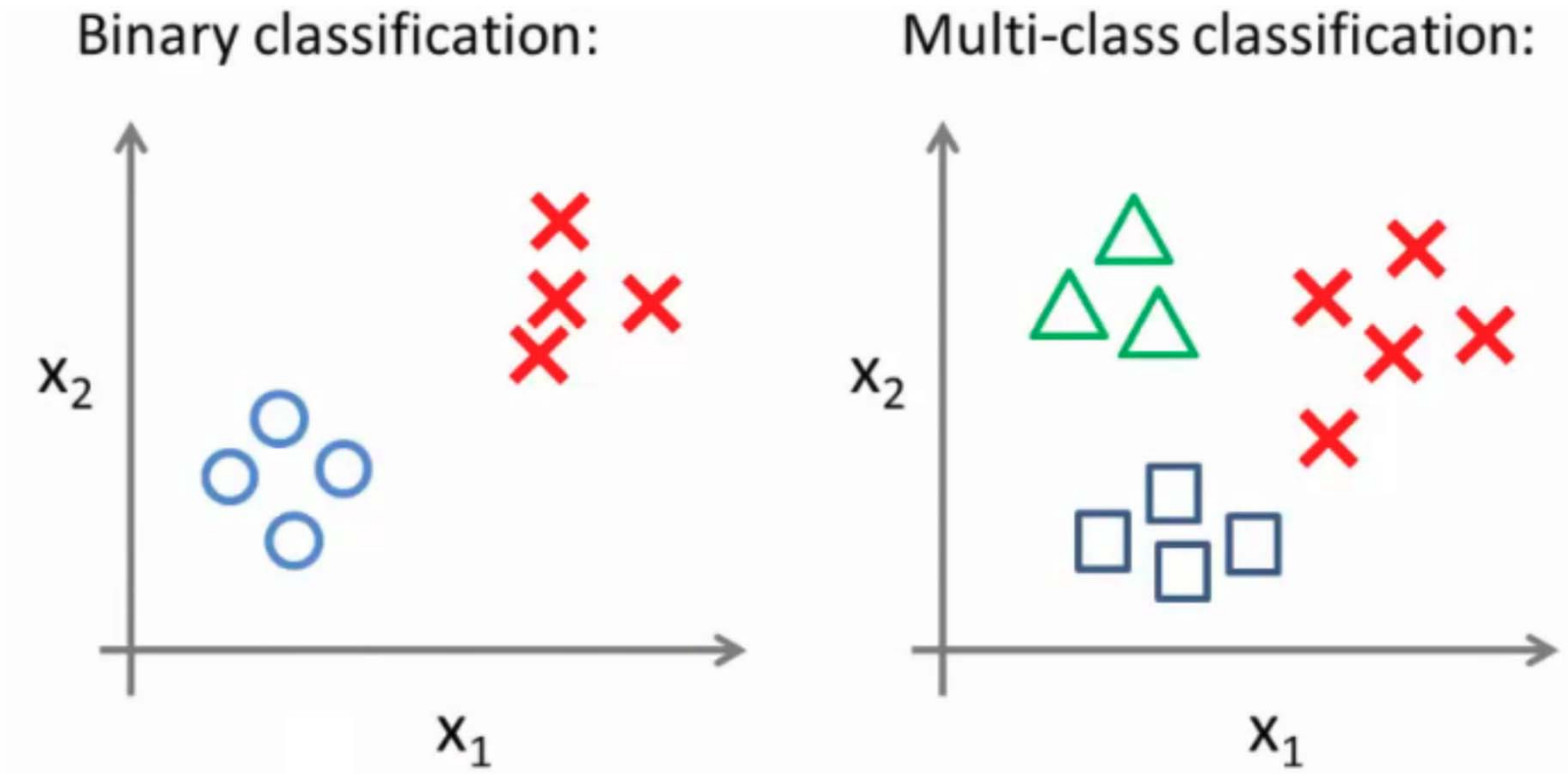
Quick bonus question

Do you have any other idea to handle imbalanced data?

i.e., one class is frequent, the other is infrequent

Multi-class logistic regression

Multi-class logistic regression



Model

- Class $C = \{c_1, c_2, \dots c_K\}$
- The probability of class k

$$p_\theta(y = c_k|\boldsymbol{x}) = \frac{e^{\boldsymbol{\theta}'_k \boldsymbol{x}}}{\sum_{k=1}^{K} e^{\boldsymbol{\theta}'_k \boldsymbol{x}}} \text{ for } k = 1, \dots, K$$

(softmax function)

- Parameters
  - $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_K$
  - Can be normalized to be K-1 groups of parameters
    - $p_\theta(y = c_k|\boldsymbol{x}) = \frac{e^{\boldsymbol{\theta}'_k \boldsymbol{x}}}{\mathbf{1} + \sum_{k=1}^{K-1} e^{\boldsymbol{\theta}'_k \boldsymbol{x}}}$ for $k = 1, \dots, K-1$
    - $p_\theta(y = c_K|\boldsymbol{x}) = \frac{1}{\mathbf{1} + \sum_{k=1}^{K-1} e^{\boldsymbol{\theta}'_k \boldsymbol{x}}}$

Strategy

- Cross entropy

$$L(y, \boldsymbol{x}, p_\theta) = -\sum_{k=1}^{K} 1_{y=c_k} \log p_\theta(c_k|\boldsymbol{x})$$

  - Empirical error

$$\hat{R}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K) = -\sum_{i=1}^{N} \log p_\theta(y_i|\boldsymbol{x})$$

- Likelihood

$$\mathcal{L}(y, \boldsymbol{x}, p_\theta) = \prod_{i=1}^{N} p_\theta(y_i|\boldsymbol{x})$$

  - Take log:

$$\log \mathcal{L}(y, \boldsymbol{x}, p_\theta) = \sum_{i=1}^{N} \log p_\theta(y_i|\boldsymbol{x})$$

Minimize empirical error with cross entropy loss

$\Updownarrow$

Maximize the log likelihood

Algorithm

- Gradient

$$\frac{\partial}{\partial \boldsymbol{\theta}_k} \log p_\theta(y = c_k | \boldsymbol{x}) = \frac{\partial}{\partial \boldsymbol{\theta}_k} \log \frac{e^{\boldsymbol{\theta}'_k \boldsymbol{x}}}{\sum_{k=1}^{K} e^{\boldsymbol{\theta}'_k \boldsymbol{x}}}$$

$$= \frac{\partial}{\partial \boldsymbol{\theta}_k} \log e^{\boldsymbol{\theta}'_k \boldsymbol{x}} - \frac{\partial}{\partial \boldsymbol{\theta}_k} \log \sum_{k=1}^{K} e^{\boldsymbol{\theta}'_k \boldsymbol{x}}$$

$$= \boldsymbol{x} - \frac{e^{\boldsymbol{\theta}'_k \boldsymbol{x}} \boldsymbol{x}}{\sum_{k=1}^{K} e^{\boldsymbol{\theta}'_k \boldsymbol{x}}} = \boldsymbol{x}(1 - p_\theta(y = c_k | \boldsymbol{x}))$$

Application

Diabetes Care

Service Quality vs. Efficiency

- Key trade-off in many service systems
  - Healthcare
- Improve operational efficiency without sacrificing quality
- How to assess quality?
  - Quality of decisions vs. outcomes (risk)
  - Critical decisions are often made by people with expert knowledge like physicians
- How to incorporate quality into operations management models?
  - Need quantitative and objective assessment of quality

Healthcare Quality Assessment

- Importance: Critical in improving care quality and efficiency of healthcare operations
  - Timely intervention to revert poor quality care
  - Capacity shortage in healthcare systems
- Challenge: No single set of guidelines for defining quality of healthcare
- How?
  - Health professionals are experts in quality of care assessment

Experts Assessment

- Healthcare quality assessment through expert opinions
  - Expert physicians can evaluate quality by examining a patient's records
  - This process is time consuming and inefficient
  - Experts are limited by memory and time
  - They cannot assess quality for millions of patients
- Similar practice in other industries
  - Accreditation in education, manufacturing, etc.

Replicating Expert Assessment

- Can we develop analytical tools that replicate expert assessment?
- Learn from expert human judgment
  - Develop a model, interpret results, and adjust the model
- Make predictions/evaluations on a large scale

Claims Data

Medical Claims

Diagnosis, Procedures, Doctor/Hospital, Cost

Pharmacy Claims

Drug, Quantity, Doctor, Medication Cost

- Electronically available
- Standardized

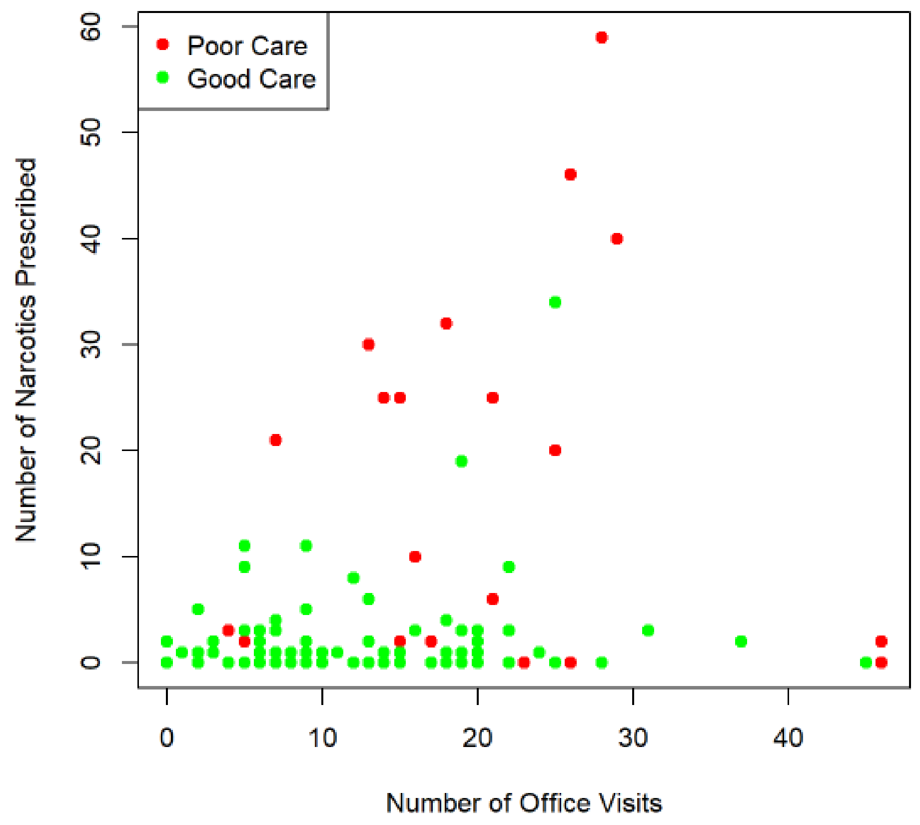
- Not 100% accurate
- Under-reporting is common
- Claims for hospital visits can be vague

Building a model

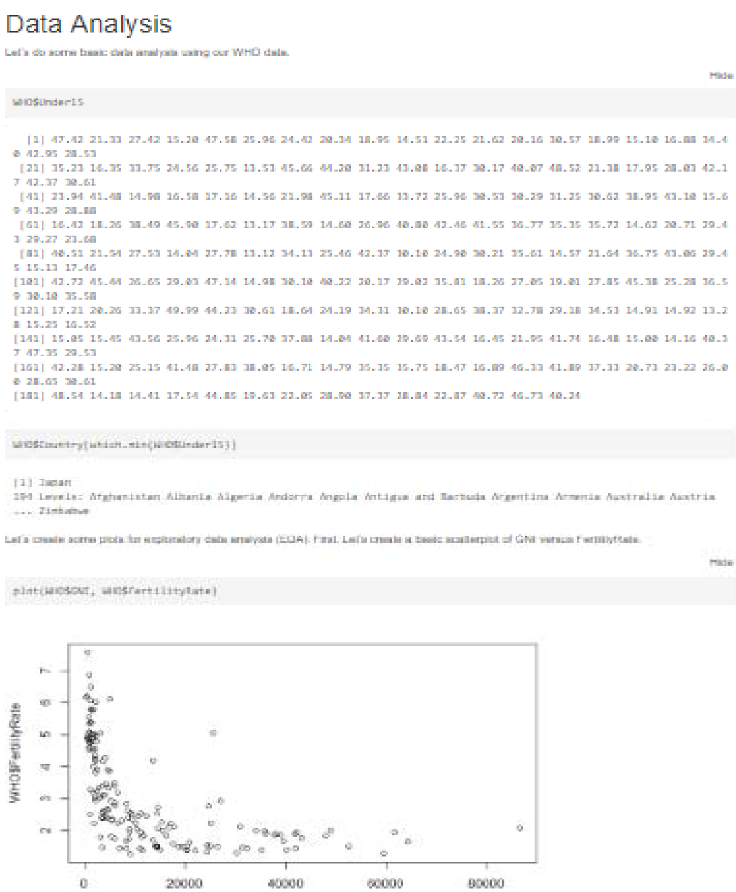
- We use a **logistic regression**
  - Predicts an outcome variable, or *dependent/response variable*
  - Using a set of *independent/explanatory variables*
- Dependent variable: Poor care or not
  - is equal to 1 if the patient had poor care, and equal to 0 if the patient had good care
- Independent variables:
  - Number of Office Visits (OfficeVisits)
  - Number of Narcotics Prescribed (Narcotics)
  - Etc.

Model for Healthcare Quality

- Plot of the independent variables
  - Number of Office Visits (OfficeVisits)
  - Number of Narcotics Prescribed (Narcotics)
- Red are poor care
- Green are good care
- Are these variables predictive of good care or poor care?



Let’s get our hands dirty!



Data Analysis

Let's do some basic data analysis using our WHO data.

Hide

```
WHO$Under15
```

```
  [1] 47.42 21.33 27.42 15.20 47.58 25.96 24.42 20.34 18.95 14.51 22.25 21.62 20.16 30.57 18.99 15.10 16.88 34.4
0 42.95 28.53
 [21] 35.23 16.35 33.75 24.56 25.75 13.53 45.66 44.20 31.23 43.08 16.37 30.17 40.07 48.52 21.38 17.95 28.03 42.1
7 42.37 30.61
 [41] 23.94 41.48 14.98 16.58 17.16 14.56 21.98 45.11 17.66 33.72 25.96 30.53 30.29 31.25 30.62 38.95 43.10 15.6
9 43.29 28.88
 [61] 16.42 18.26 38.49 45.90 17.62 13.17 38.59 14.60 26.96 40.80 42.46 41.55 36.77 35.35 35.72 14.62 20.71 29.4
3 29.27 23.68
 [81] 40.51 21.54 27.53 14.04 27.78 13.12 34.13 25.46 42.37 30.10 24.90 30.21 35.61 14.57 21.64 36.75 43.06 29.4
5 15.13 17.46
[101] 42.72 45.44 26.65 29.03 47.14 14.98 30.10 40.22 20.17 29.02 35.81 18.26 27.05 19.01 27.85 45.38 25.28 36.5
9 30.10 35.58
[121] 17.21 20.26 33.37 49.99 44.23 30.61 18.64 24.19 34.31 30.10 28.65 38.37 32.78 29.18 34.53 14.91 14.92 13.2
8 15.25 16.52
[141] 15.05 15.45 43.56 25.96 24.31 25.70 37.88 14.04 41.60 29.69 43.54 16.45 21.95 41.74 16.48 15.00 14.16 40.3
7 47.35 29.53
[161] 42.28 15.20 25.15 41.48 27.83 38.05 16.71 14.79 35.35 35.75 18.47 16.89 46.33 41.89 37.33 20.73 23.22 26.0
0 28.65 30.61
[181] 40.54 14.18 14.41 17.54 44.85 19.63 22.05 28.90 37.37 28.84 22.87 40.72 46.73 40.24
```

```
WHO$Country[which.min(WHO$Under15)]
```

```
[1] Japan
194 Levels: Afghanistan Albania Algeria Andorra Angola Antigua and Barbuda Argentina Armenia Australia Austria
... Zimbabwe
```

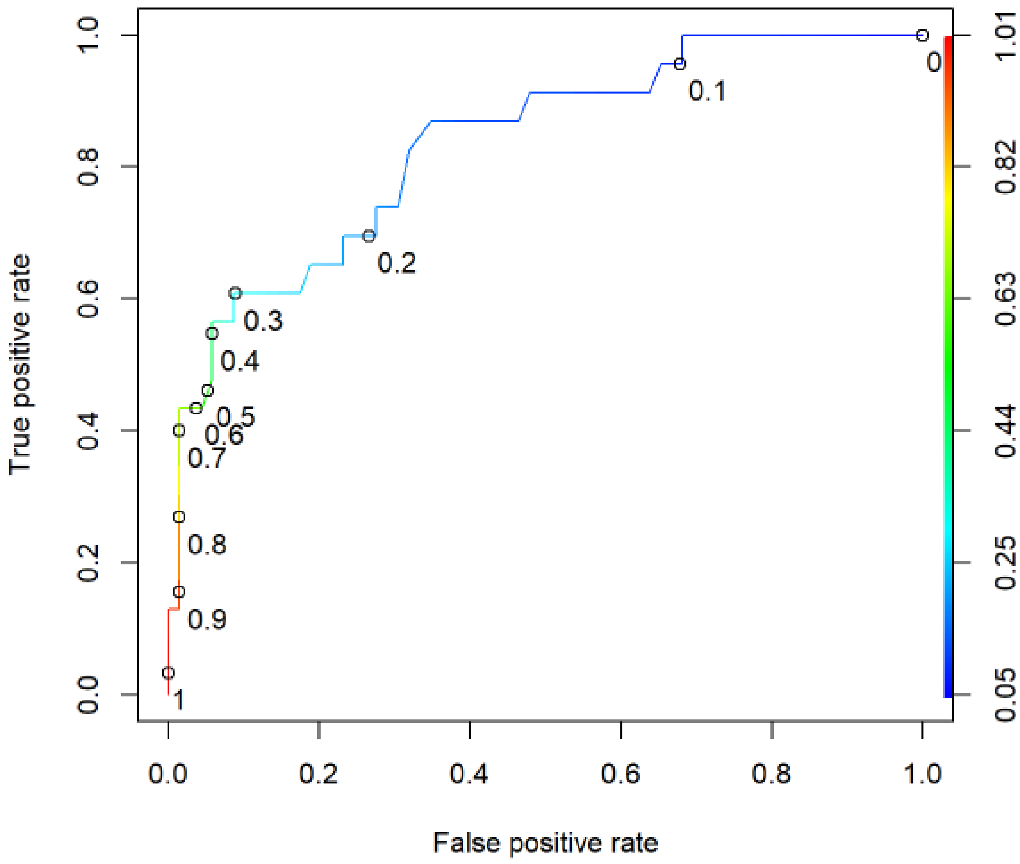
Let's create some plots for exploratory data analysis (EDA). First, Let's create a basic scatterplot of GNI versus FertilityRate.

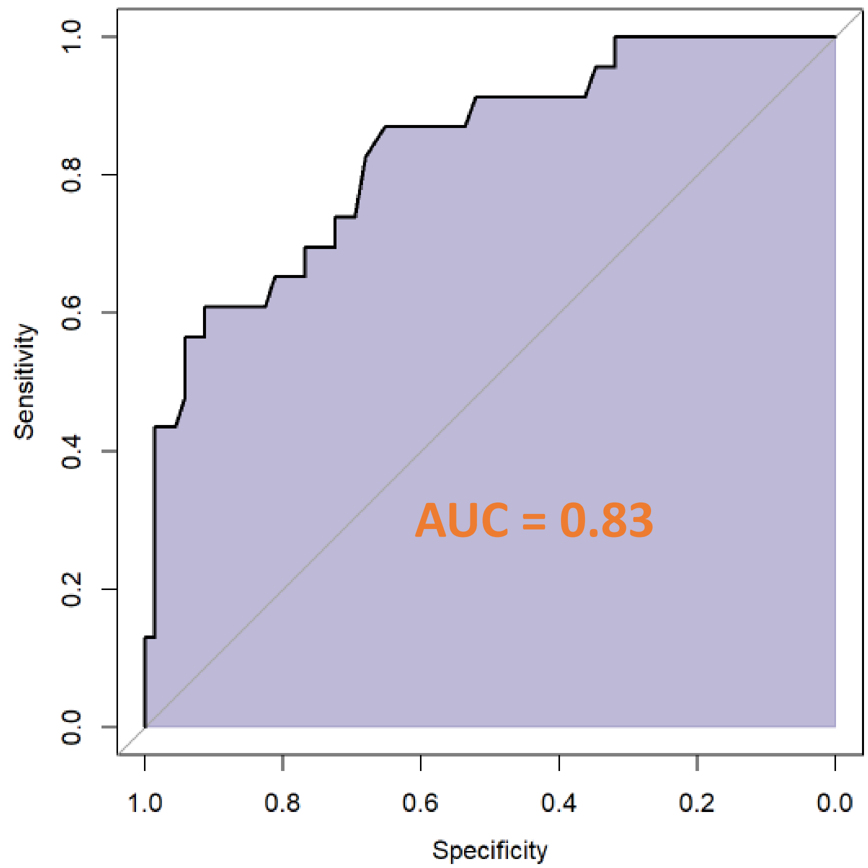
Hide

```
plot(WHO$GNI, WHO$FertilityRate)
```


Application: Diabetes Care

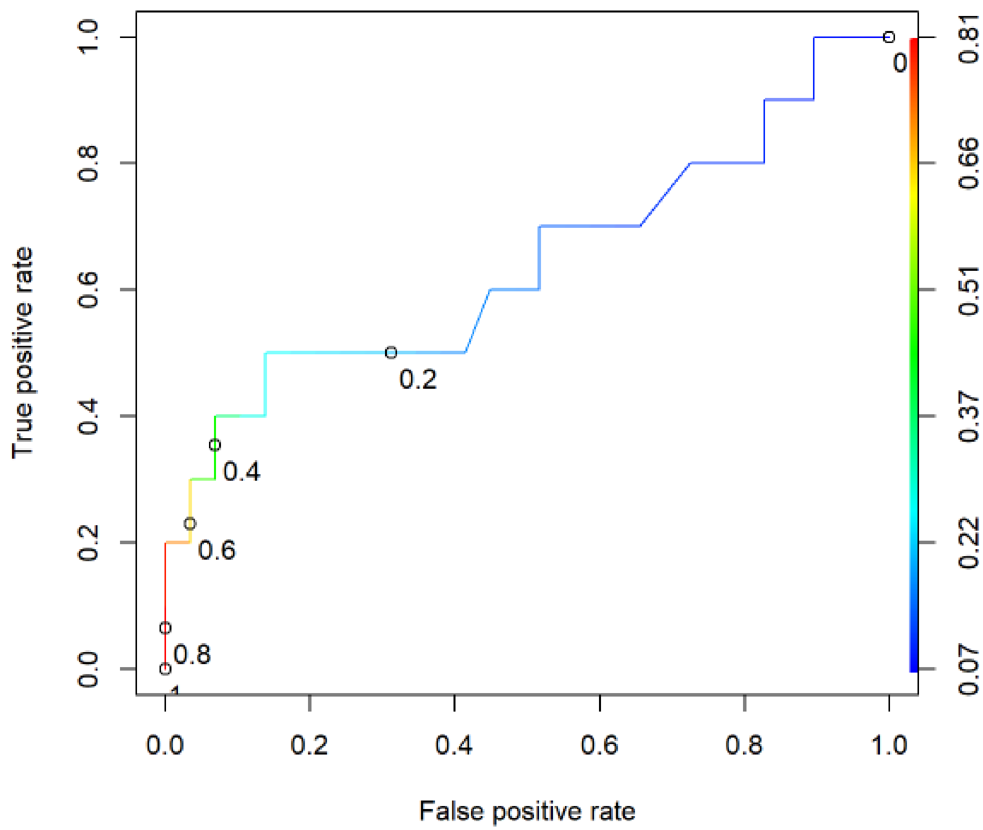
ROC and AUC





Making Predictions

- Out-of-sample ROC
- Out-of-sample AUC = 0.64



Lecture 4 wrap-up

- ✓ Classification problem
- ✓ Logistic regression method
  - ✓ Model
  - ✓ Strategy
  - ✓ Algorithm
- ✓ Prediction & evaluation
- ✓ Multi-class logistic regression
- ✓ Application: diabetes care

Next lecture

- Supervised learning
  - Linear regression
  - Logistic regression
  - SVM and kernel
  - Tree models
- Unsupervised learning
  - Clustering
  - PCA
  - EM
- Deep learning
  - Neural networks
  - Convolutional NN
  - Recurrent NN
- Reinforcement learning
  - MDP
  - ADP
  - Deep Q-Network

Questions?

Shan Wang (王杉)
https://wangshan731.github.io/