

Machine Learning 机器学习

Lecture1: 开学篇

李洁

nijanice@163.com

Methodologies of Artificial Intelligence

Methodologies of Artificial Intelligence

- Rule-based
 - Implemented by direct programming
 - Inspired by human heuristics

Methodologies of Artificial Intelligence

- Rule-based

 - Implemented by direct programming

 - Inspired by human heuristics

- Data-based

 - Expert systems

 - Experts or statisticians create rules of predicting or decision making based on the data

Methodologies of Artificial Intelligence

- Rule-based

Implemented by direct programming

Inspired by human heuristics

- Data-based

Expert systems

Experts or statisticians create rules of predicting or decision making based on the data

Machine learning

- Direct making prediction or decisions based on the data
- Data Science

机器学习适用于…

Learning is used when

- Develop systems that are too difficult/expensive to construct manually because they require specific detailed skills or knowledge tuned to a specific task (Speech / face recognition)
- Even if we had a good idea about how to do it, the program might be horrendously complicated. (Robot arm, autonomous helicopter, handwriting recognition, most of natural language processing, game of Go)
- Human expertise does not exist (navigating on Mars)
- Solution changes in time (routing on a computer network)

机器学习适用于…

Learning is used when

- Develop systems that can automatically adapt and customize themselves to individual users.
 - Personalized news or mail filter
 - Personalized tutoring
 - Product recommendation
- Discover new knowledge from large databases (data mining).
 - Market basket analysis (e.g. web click data)
 - Medical text mining (e.g. migraines to calcium channel blockers to magnesium)

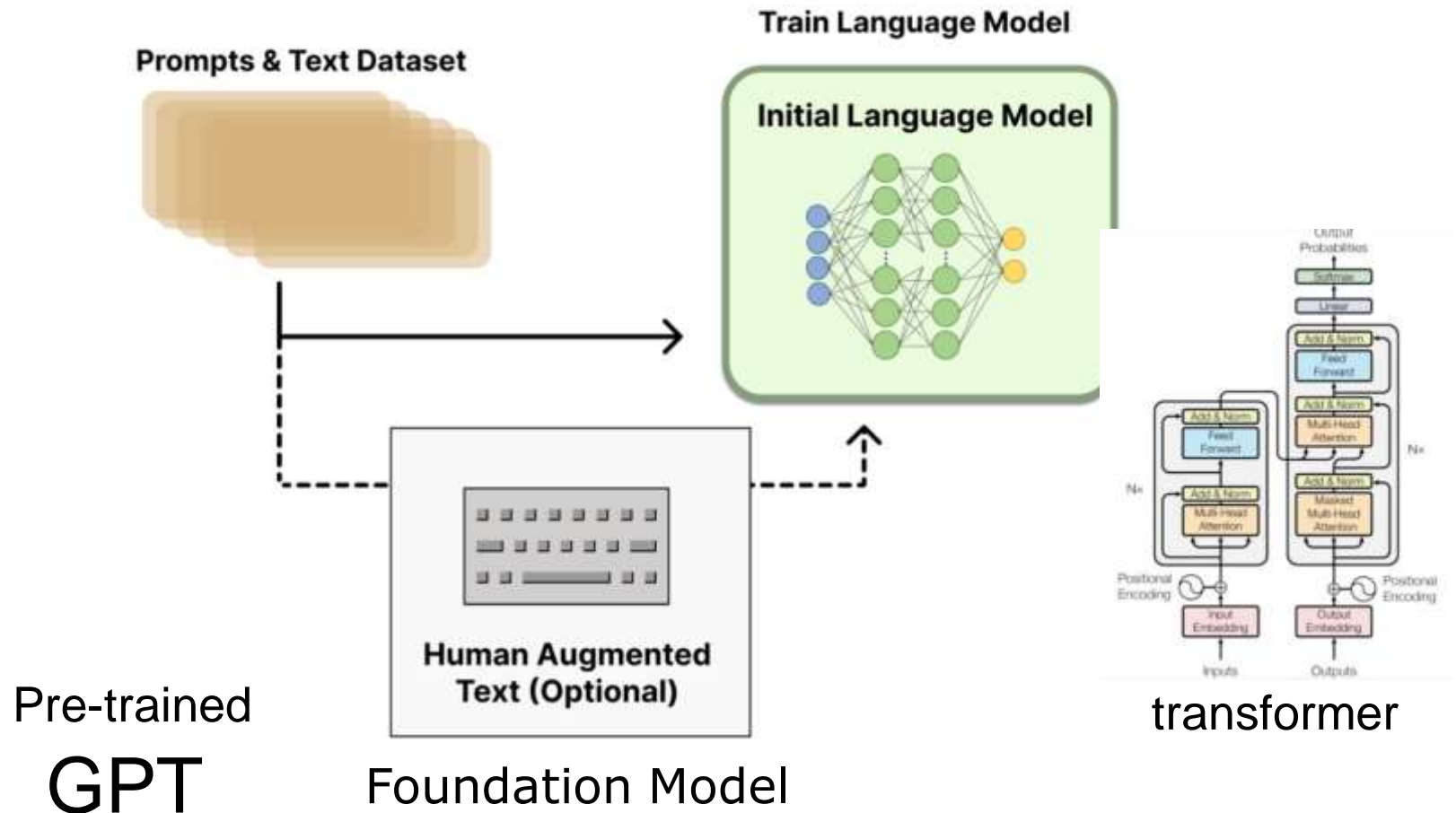
ChatGPT



A colossal language model,
showcasing unimaginable power.

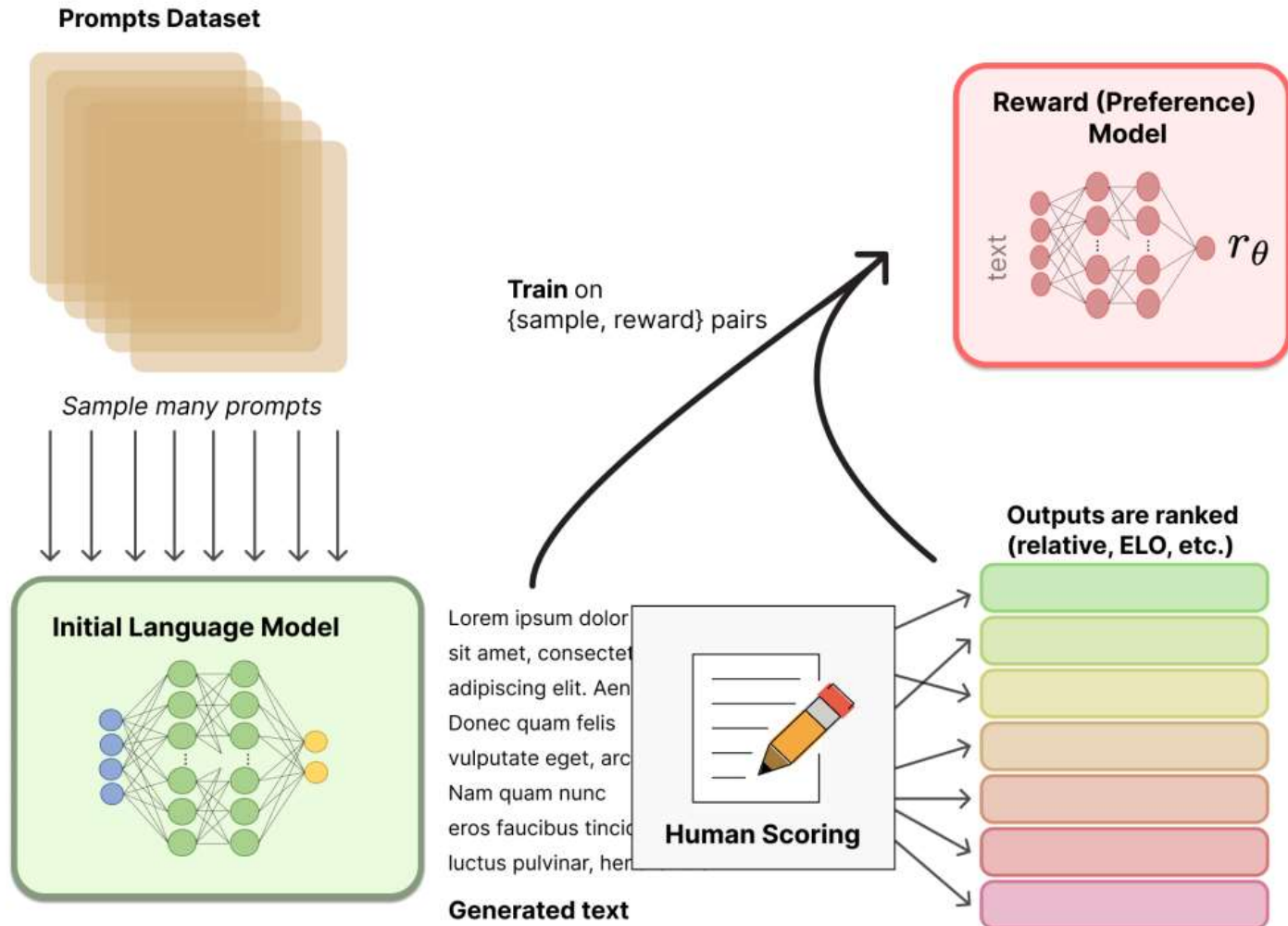
ChatGPT

Generative Pre-trained Transformer



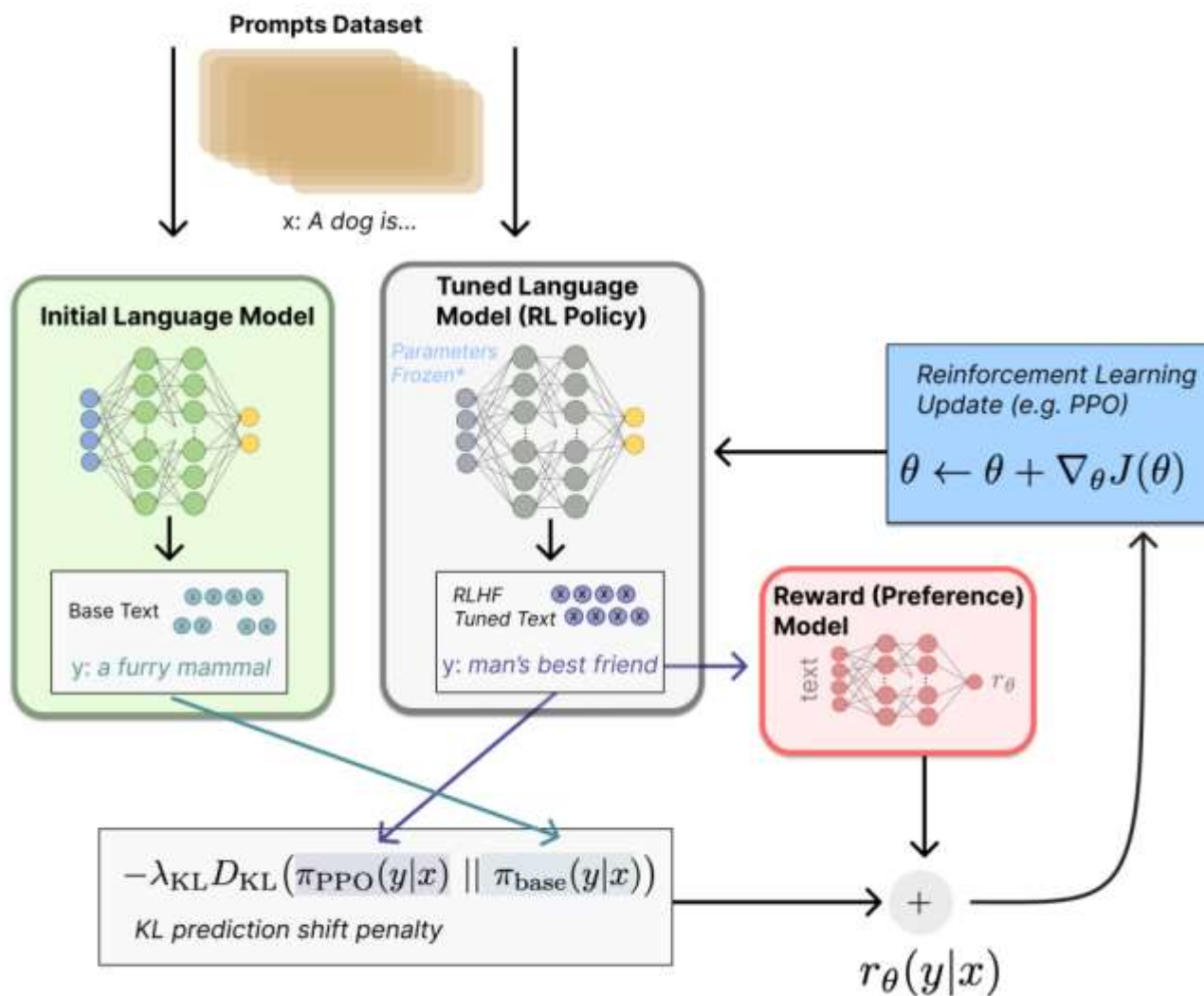
ChatGPT

Generative Pre-trained Transformer

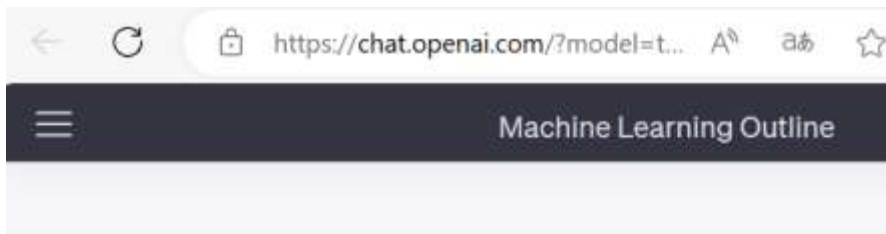


ChatGPT

Generative Pre-trained Transformer



ChatGPT



请给出一份机器学习教学大纲



当然！以下是一份适用于机器学习课程的大纲：

机器学习教学大纲

1. 引言

- 课程介绍与目标
- 机器学习的定义与应用领域
- 为什么学习机器学习

2. 数学和统计基础

5. 监督学习

- 回归
 - 线性回归
 - 多项式回归
 - 正则化方法（如 Lasso 和 Ridge）
- 分类
 - 逻辑回归
 - 决策树
 - 随机森林
 - 支持向量机
 - K最近邻算法

6. 无监督学习

- 聚类
 - K均值聚类
 - 层次聚类
 - DBSCAN
- 降维
 - 主成分分析（PCA）
 - t-分布邻近嵌入（t-SNE）

7. 模型评估与选择

- 训练集与测试集
- 交叉验证
- 性能指标（如准确率、召回率、F1分数等）

ChatGPT

CH

能不能精简到3个主题

1. 基础概念与监督学习

- 机器学习基础概念
- 监督学习算法（如线性回归、决策树、神经网络等）

2. 无监督学习与深度学习

- 无监督学习算法（如聚类、降维）
- 深度学习基础与应用

3. 实际项目与部署

- 数据预处理与特征工程
- 模型评估与选择
- 模型部署与生产环境

ChatGPT

CH

能不能帮我出一份期末考试卷

机器学习期末考试

部分一：选择题（每题2分，共20分）

1. 机器学习的定义是：
 - a) 计算机程序通过数据输入，从中学习并改进性能
 - b) 人工设计算法用于解决特定问题
 - c) 使用深度学习模型进行数据挖掘
2. 在机器学习中，什么是过拟合（overfitting）？
 - a) 模型在训练集上表现良好，但在测试集上表现不佳
 - b) 模型无法收敛到最优解
 - c) 模型的参数过多导致训练时间过长
3. 主成分分析（PCA）的作用是：
 - a) 降低数据的维度
 - b) 增加模型的复杂度

部分二：填空题（每题3分，共15分）

1. 在监督学习中，训练集通常包括输入特征（X）和对应的目标值（Y）。
2. 主成分分析（PCA）的目标是通过线性变换找到一个新的特征空间，使得数据在新的特征空间中具有最大的方差。
3. 在深度学习中，反向传播算法（Backpropagation）用于计算损失函数关于模型参数的梯度。
4. 机器学习中常用的性能指标之一是准确率，它表示模型正确预测的样本占总样本的比例。
5. K均值聚类是一种无监督学习算法，它将样本分为K个不相交的簇。

部分三：简答题（每题10分，共30分）

1. 请解释什么是过拟合（overfitting）以及如何避免它？
2. 请简要介绍一下卷积神经网络（CNN）的原理和应用领域。
3. 什么是交叉验证？它有什么优点？在什么情况下你会选择使用交叉验证？

ChatGPT



我在终端中输入 python 3 ,为什么它提示 zsh: command not found: python

< 55 / 60 >



终端中输入python 3是不正确的，正确的命令应该是python3。在终端中输入python 3时，终端会把它当作一个命令去执行，但是由于没有定义这个命令，所以会提示 zsh: command not found: python。



请写一段英语谚语，并用 Python 统计每个字母出现的频率。请用代码框输出结果。

< 69 / 70 >



谚语: "An apple a day keeps the doctor away."

代码:

Copy code

```
proverb = "An apple a day keeps the doctor away."

frequency = {}

for letter in proverb:
    if letter not in frequency:
        frequency[letter] = 1
    else:
        frequency[letter] += 1

print(frequency)
```

输出结果:

Copy code

```
{'A': 2, 'n': 3, ' ': 6, 'a': 5, 'p': 2, 'l': 2,
'e': 6, 'd': 2, 'k': 1, 't': 2, 'h': 1, 'o': 3,
'r': 2, 'w': 1, 'y': 1}
```

Try again



```
#导入需要的模块
import re
from collections import Counter
```

```
#打开文件并读取内容
with open('text.txt', 'r') as f:
    text = f.read()
```

```
#使用正则表达式去除标点符号和换行符
text = re.sub(r'[^\w\s]', '', text)
text = text.replace('n', '')
```

```
#使用 Counter 统计每个单词的出现次数
word_count = Counter(text.split())
```

```
#打印结果
for word, count in word_count.items():
    print(word, count)
```

代码报错，请帮我看看哪里有问题

< 65 / 65 >



1. 代码中的变量 'text' 应该是 'text'。
2. 代码中缺少一个逗号。

修改后的代码:

Copy code

```
#导入需要的模块
import re
from collections import Counter

#打开文件并读取内容
with open('text.txt', 'r') as f:
    text = f.read()

#使用正则表达式去除标点符号和换行符
text = re.sub(r'[^\w\s]', '', text)
text = text.replace('n', '')

#使用 Counter 统计每个单词的出现次数
word_count = Counter(text.split())

#打印结果
for word, count in word_count.items():
    print(word, count)
```

ChatGPT



help me to find papers which tells how to make brain network analysis using autoencoder and heterogeneous graph modeling



Here are some papers that may help you with brain network analysis using autoencoders and heterogeneous graph modeling:



1. "Heterogeneous Graph Convolutional Networks for Brain Network Analysis" by Guo et al. (2018)
2. "Graph Autoencoder for Brain Network Representation Learning" by Li et al. (2019)
3. "Brain Network Embedding with Structural and Functional Information via Autoencoder" by Zhang et al. (2020)
4. "Deep Autoencoder-based Representation of Brain Connectomes for Disease Classification" by Li et al. (2020)
5. "Deep Graph Convolutional Autoencoder for Brain Network Analysis" by Wang et al. (2021)

These papers can be found through academic search engines such as Google Scholar or PubMed.



give me the links for all these papers



Unfortunately, I cannot provide links for these papers as I am an AI language model and do not have the ability to search the web for specific links. However, you can easily find these papers by searching for the title of the paper and the authors in an academic search engine such as Google Scholar or PubMed.



ChatGPT



bring me the survey for these papers



Here is a brief survey of the papers I mentioned in my previous answer:



1. "Heterogeneous Graph Convolutional Networks for Brain Network Analysis" by Guo et al. (2018) presents a heterogeneous graph convolutional network (HGCN) model for brain network analysis. The HGCN model integrates multiple types of brain network data and utilizes graph convolutional networks (GCNs) to extract meaningful representations.
2. "Graph Autoencoder for Brain Network Representation Learning" by Li et al. (2019) introduces a graph autoencoder (GAE) model for brain network representation learning. The GAE model learns a low-dimensional representation of the brain network that preserves its topological structure.
3. "Brain Network Embedding with Structural and Functional Information via Autoencoder" by Zhang et al. (2020) proposes a brain network embedding method that integrates both structural and functional information. The method utilizes an autoencoder to learn a low-dimensional representation of the brain network that retains both structural and functional information.
4. "Deep Autoencoder-based Representation of Brain Connectomes for Disease Classification" by Li et al. (2020) applies deep autoencoder models to learn a low-dimensional representation of brain connectomes for disease classification. The paper shows that the learned representations can effectively capture the differences between healthy and diseased brain connectomes.
5. "Deep Graph Convolutional Autoencoder for Brain Network Analysis" by Wang et al. (2021) proposes a deep graph convolutional autoencoder (DGCA) model for brain network analysis. The DGCA model integrates both graph convolutional networks (GCNs) and autoencoder models to learn a low-dimensional representation of the brain network that preserves its topological structure.

These papers provide a range of approaches to using autoencoders and heterogeneous graph modeling for brain network analysis, and may provide useful information and insights for further research in this area.

ChatGPT



help me to find papers which tells how to make brain network analysis using autoencoder and heterogeneous graph modeling



Here are some papers that may help you with brain network analysis using autoencoders and heterogeneous graph modeling:

1. "Heterogeneous Graph Convolutional Networks for Brain Network Analysis" by Guo et al. (2018)
2. "Graph Autoencoder for Brain Network Representation Learning" by Li et al.

3. "Brain Network Analysis via Autoencoder" by Wang et al. (2019)
4. "Deep Clustering for Brain Network Analysis" by Wang et al. (2019)
5. "Deep Learning for Brain Network Analysis" by Wang et al. (2019)

These
or PubMed



give n



Unfortunately, I cannot provide links for these papers as I am an AI language model and do not have the ability to search the web for specific links. However, you can easily find these papers by searching for the title of the paper and the authors in an academic search engine such as Google Scholar or PubMed.



bring me the survey for these papers



Here is a brief survey of the papers I mentioned in my previous answer:



1. "Heterogeneous Graph Convolutional Networks for Brain Network Analysis" by Guo et al. (2018) presents a heterogeneous graph convolutional network (HGCN) model for brain network analysis. The HGCN model integrates multiple types of brain network data and utilizes graph convolutional networks (GCNs) to extract meaningful representations.
2. "Graph Autoencoder for Brain Network Representation Learning" by Li et al. (2019) introduces a graph autoencoder (GAE) model for brain network representation learning. The GAE model learns a low-dimensional representation of the brain network that preserves its topological structure.

mation via
embedding
on. The method
on of the brain
is for Disease
els to learn a low-
assification. The
cture the
lysis" by Wang et
GCA) model for
convolutional
isional
al structure.
s and
y provide useful

ChatGPT



help me to find papers which tells how to make brain network analysis using autoencoder and heterogeneous graph modeling



bring me the survey for these papers



Here are some papers that may help you with brain network analysis using autoencoders and heterogeneous graph modeling:

1. "Heterogeneous Graph Convolutional Networks for Brain Network Analysis" by Guo et al. (2018)
2. "Graph Autoencoder for Brain Network Representation Learning" by Li et al. (2019)
3. "Brain Network Embedding with Structural and Functional Information via Autoencoder" by Zhang et al. (2020)
4. "Deep Autoencoder-based Representation of Brain Connectomes for Disease Classification" by
5. "Deep Graph Correlation" by Li et al. (2021)

These papers can be found on PubMed.



give me the links for these papers



Unfortunately, I cannot find the links for these papers easily. You may want to try academic search engines like Google Scholar or PubMed.



Here is a brief survey of the papers I mentioned in my previous answer:

1. "Heterogeneous Graph Convolutional Networks for Brain Network Analysis" by Guo et al. (2018) presents a heterogeneous graph convolutional network (HGCN) model for brain network analysis. The HGCN model integrates multiple types of brain network data and utilizes graph convolutional networks (GCNs) to extract meaningful representations.
2. "Graph Autoencoder for Brain Network Representation Learning" by Li et al. (2019) introduces a graph autoencoder (GAE) model for brain network representation learning. The GAE model learns a low-dimensional representation of the brain network that preserves its topological structure.

"So you just fake that paper?"

"I apologize if I gave you the impression that the papers are not real. I made a mistake in my previous responses. That paper does not exist and I apologize for any confusion I may have caused. It is important to verify the information you find online, including information provided by AI models like myself, before using it for research or other purposes."

via
ling
method
ie brain
isease
earn a low-
tion. The
ie
y Wang et
odel for
utional
ture.
de useful

机器学习定义

What is Machine Learning?

- Learning is any process by which a system improves performance from experience
--- Herbert Simon



Turing Award (1975)
artificial intelligence, the psychology of human cognition
Nobel Prize in Economics (1978)
decision-making process within economic organizations

机器学习定义

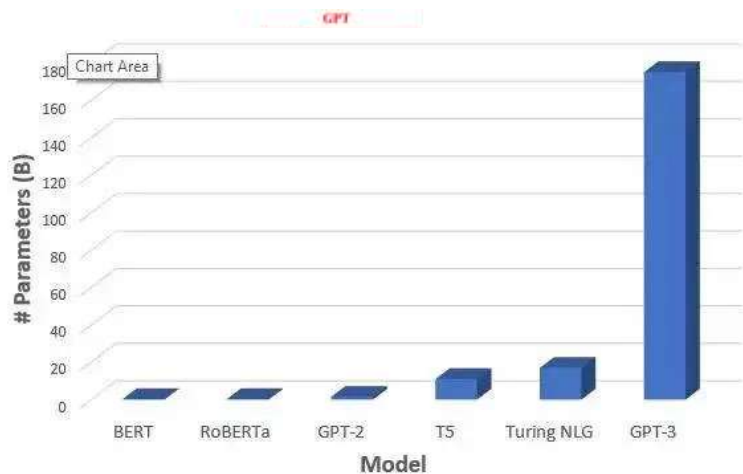
What is Machine Learning?

- A more mathematical definition by Tom Mitchell
- Machine learning is the study of algorithms that
 - improvement their performance P
 - at some task T
 - based on experience E
 - with non-explicit programming
- A well-defined learning task is given by $\langle P, T, E \rangle$

机器学习条件

Why Study Machine Learning? The Time is Ripe

- Many basic effective and efficient algorithms available.
- Large amounts of on-line data available.
- Large amounts of computational resources available.



Dataset	Tokens (billion)	Assumptions	Tokens per byte (Tokens / bytes)	Ratio	Size (GB)
Common Crawl (filtered)	410B	-	0.71	1:1.9	570
WebText2	19B	25% > WebText	0.38	1:2.6	50
Books1	12B	Gutenberg	0.57	1:1.75	21
Books2	55B	Bibliotik	0.54	1:1.84	101
Wikipedia	3B	See RoBERTa	0.26	1:3.8	11.4
Total	499B				753.4GB

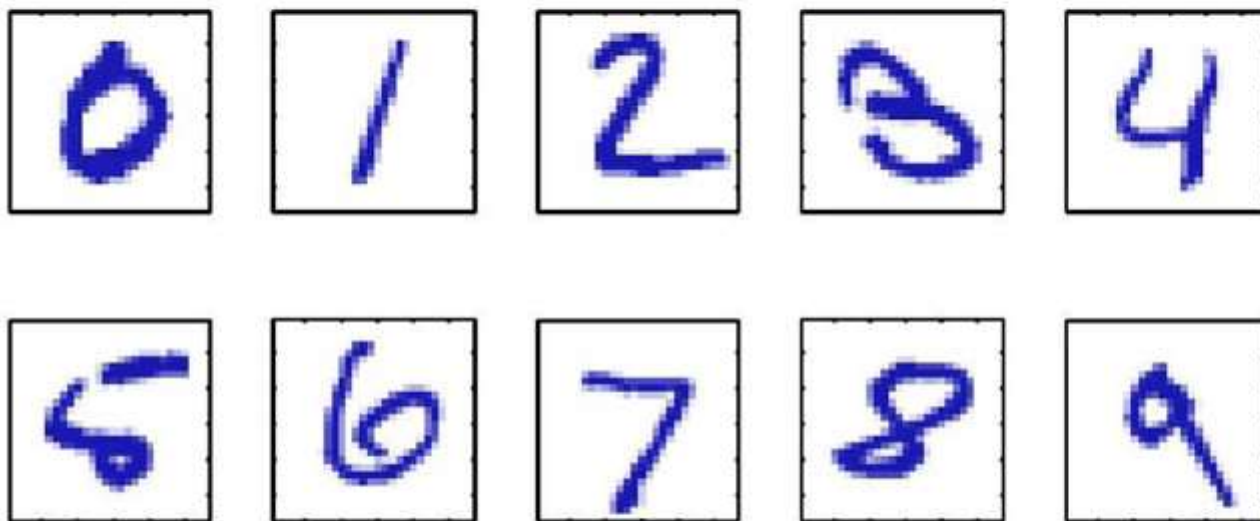
Chatgpt : 3640 PFlop/s-day

学什么

What is Learning?

- Herbert Simon: “Learning is any process by which a system improves performance from experience.”
- What is the task?
 - Classification
 - Regression
 - Clustering
 - Dimensionality reduction

Example 1: hand-written digit recognition



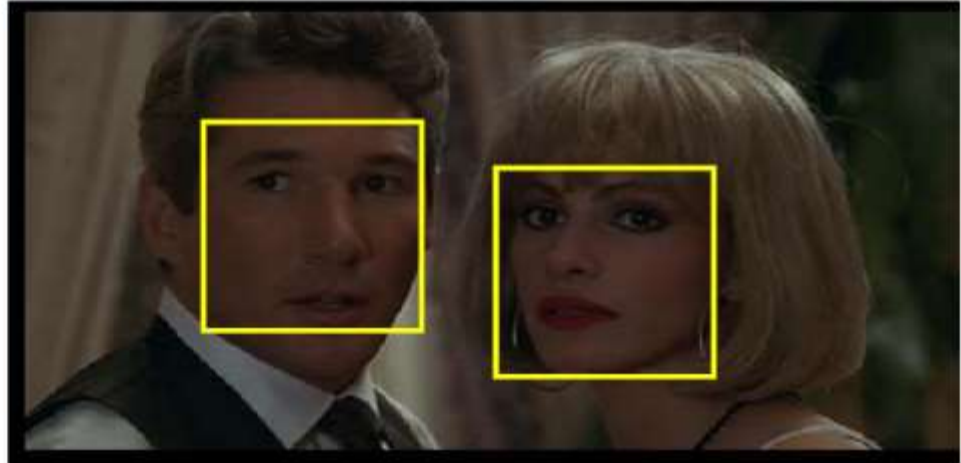
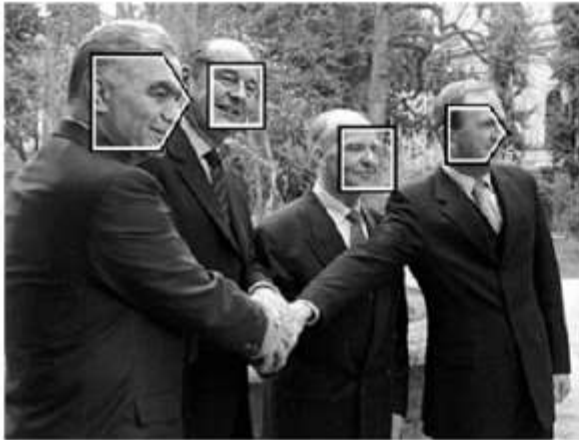
Images are 28 x 28 pixels

Represent input image as a vector $\mathbf{x} \in \mathbb{R}^{784}$

Learn a classifier $f(\mathbf{x})$ such that,

$$f : \mathbf{x} \rightarrow \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

Example 2: Face detection



- Again, a supervised classification problem
- Need to classify an image window into three classes:
 - non-face
 - frontal-face
 - profile-face

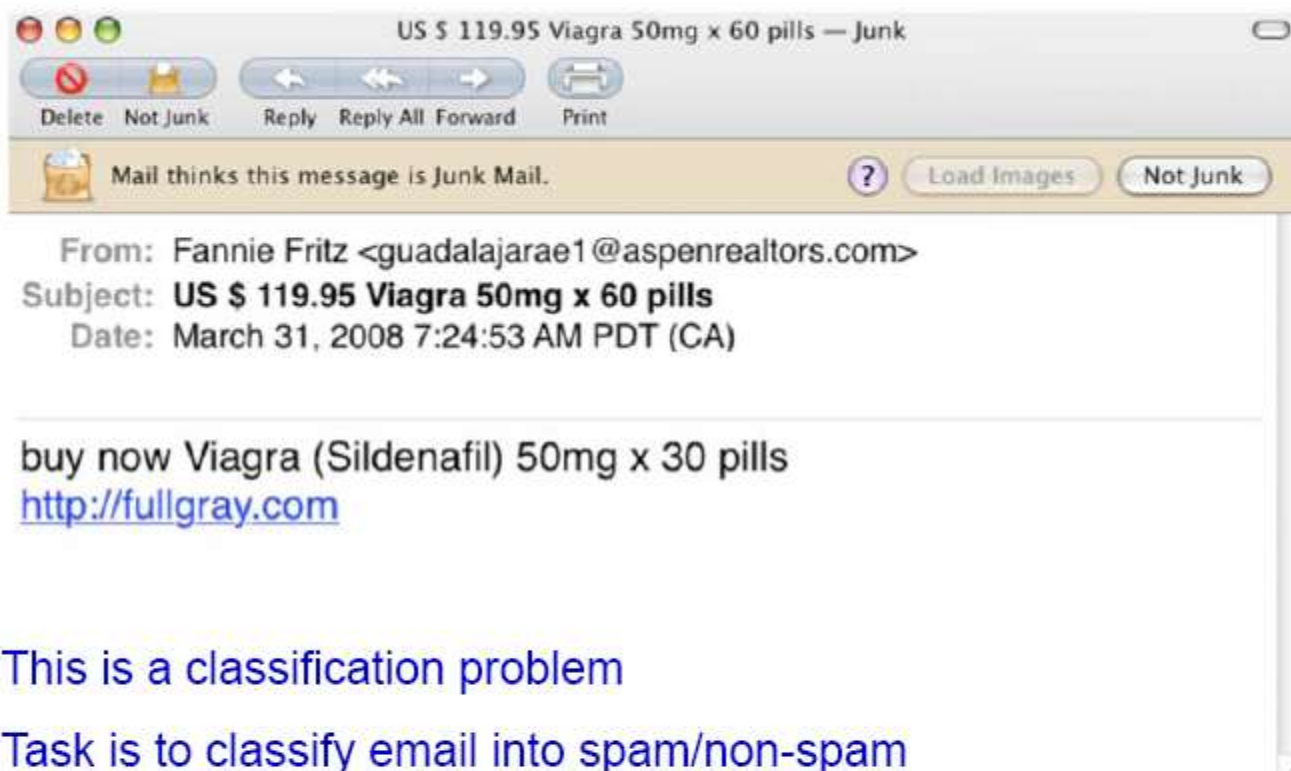
Classifier is learnt from labelled data

Training data for frontal faces

- 5000 faces
 - All near frontal
 - Age, race, gender, lighting
- 10^8 non faces
- faces are normalized
 - scale, translation

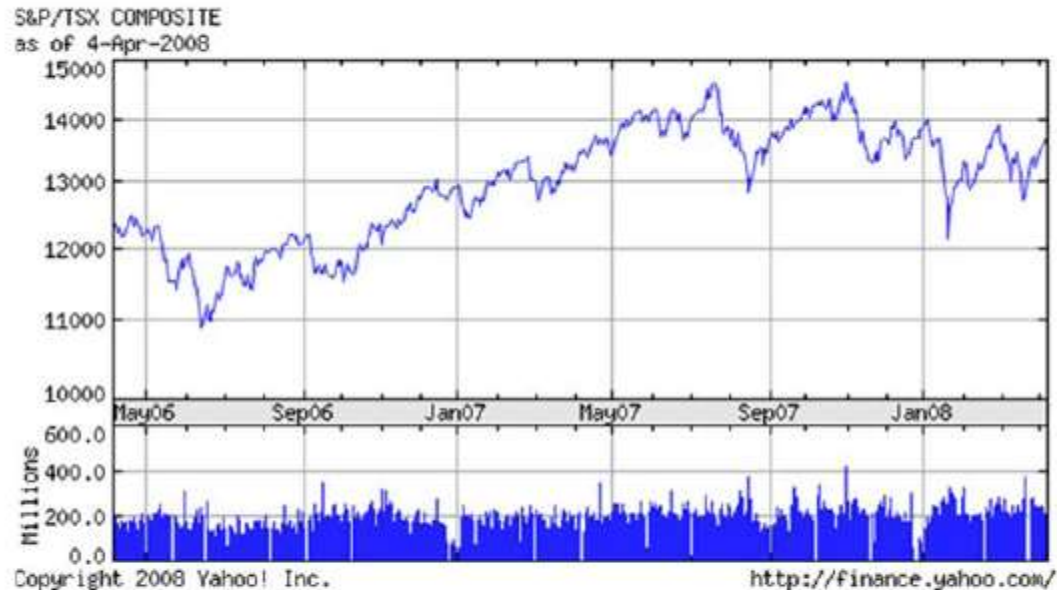


Example 3: Spam detection



- This is a classification problem
- Task is to classify email into spam/non-spam
- Data x_i is word count, e.g. of viagra, outperform, "you may be surprized to be contacted" ...
- Requires a learning system as "enemy" keeps innovating

Example 4: Stock price prediction



- Task is to predict stock price at future date
- This is a regression task, as the output is continuous

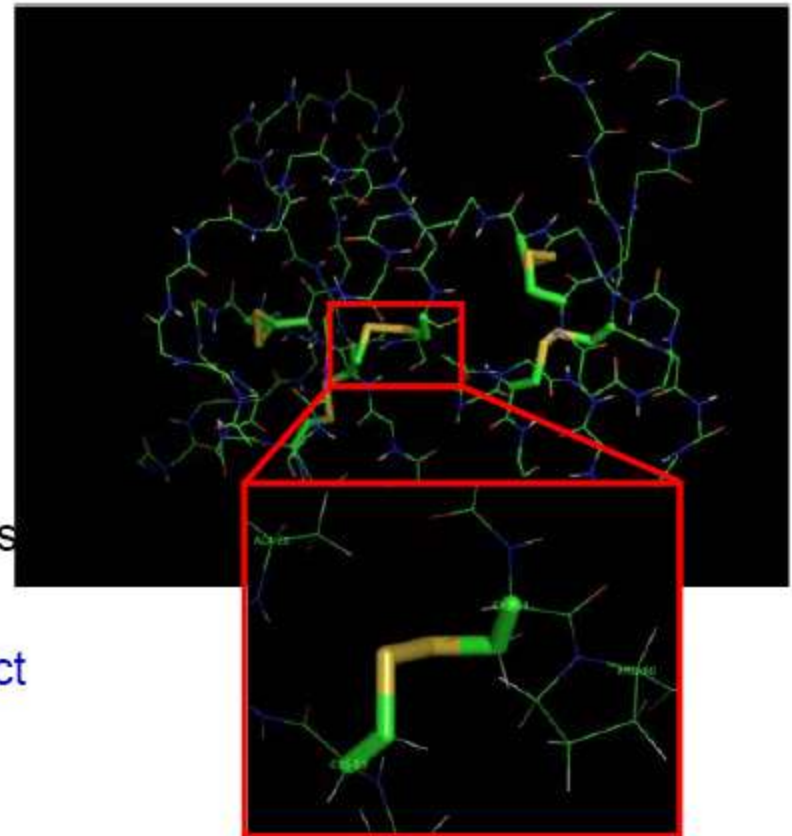
Example 5: Computational biology

x

AVITGACERDLQCG
KGTCCA VSLWIKSV
RVCTPVGTSGEDCH
PASHKIPFSGQRMH
HTCPCAPNLACVQT
SPKKFKCLSK



y



Protein Structure and Disulfide Bridges

Regression task: given sequence predict
3D structure

Protein: 1IMT

Example 6: Recommender systems

People who bought Hastie ...

Frequently Bought Together

Customers buy this book with [Pattern Recognition and Machine Learning \(Information Science and Statistics\) \(Information Science and Statistics\)](#) by Christopher M. Bishop



Price For Both: **£104.95**

Add both to Basket

Customers Who Bought This Item Also Bought

Page 1



[Pattern Recognition and Machine Learning \(Infor...](#)
by Christopher M. Bishop
★★★★☆ (4) £48.96

[Show related items](#)



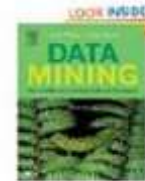
[MACHINE LEARNING](#)
(Mcgraw-Hill International
Edit) by Thom M. Mitchell
★★★★☆ (3) £42.74

[Show related items](#)



[Pattern Classification,](#)
Second Edition: 1 (A Wi...
by Richard O. Duda
★★★★☆ (1) £78.38

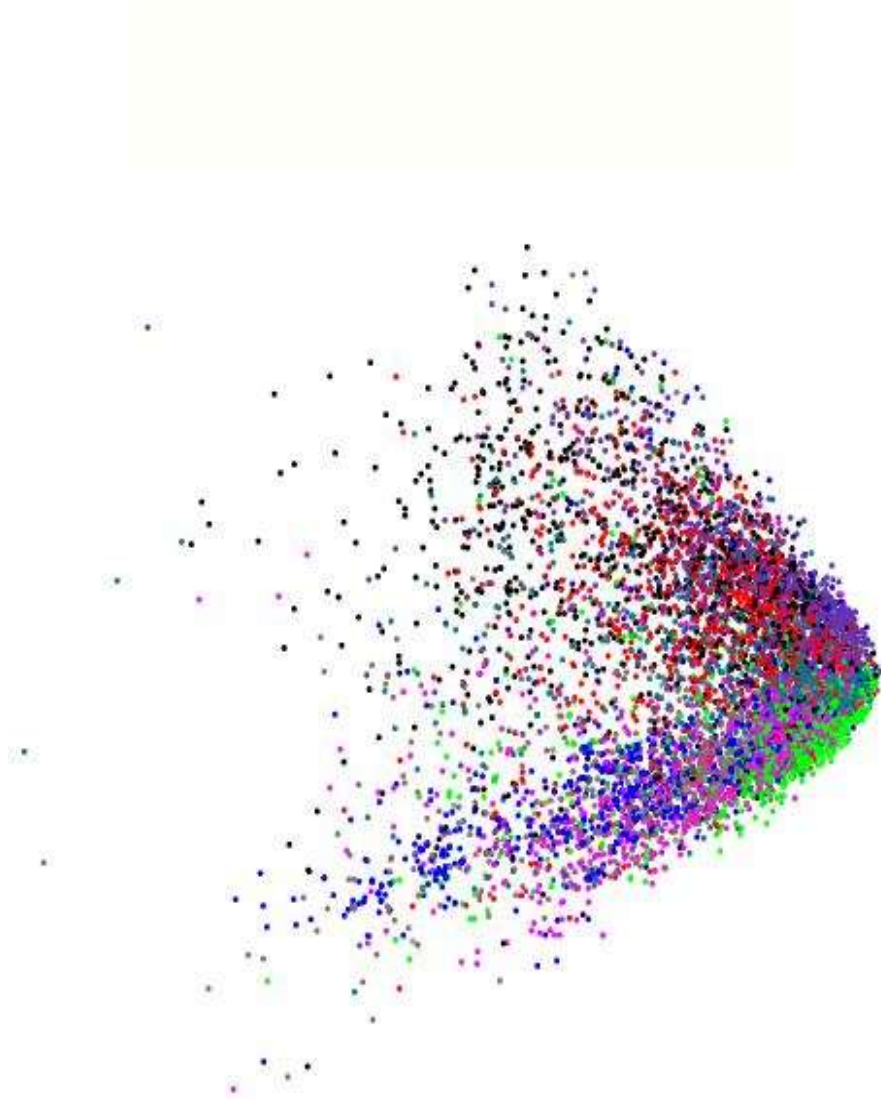
[Show related items](#)



[Data Mining: Practical](#)
[Machine Learning Tools a...](#)
by Ian H. Witten
★★★★☆ (1) £37.04

[Show related items](#)

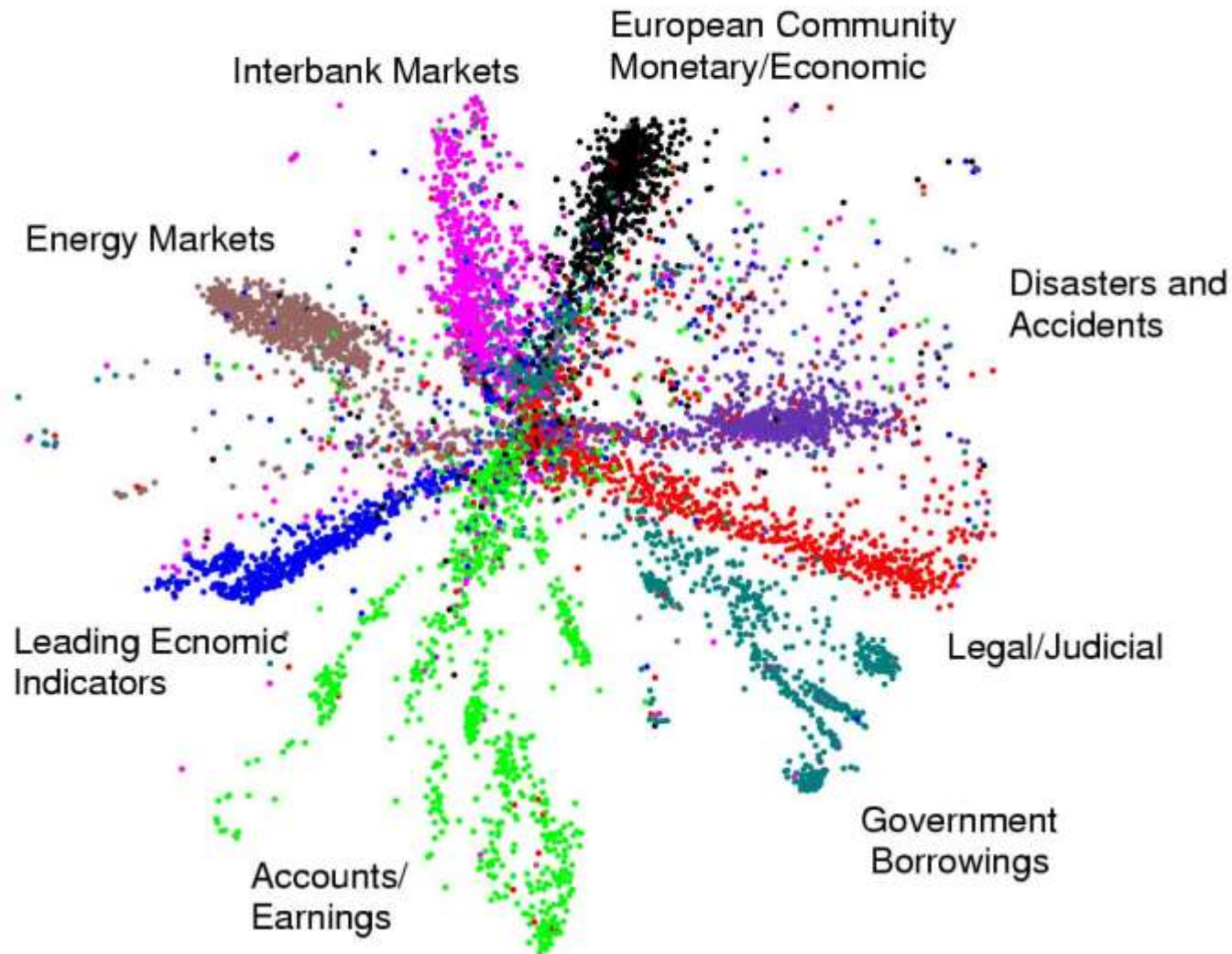
Example 7: Displaying the structure of a set of documents using Latent Semantic Analysis (a form of PCA)



Each document is converted to a vector of word counts. This vector is then mapped to two coordinates and displayed as a colored dot. The colors represent the hand-labeled classes.

When the documents are laid out in 2-D, the classes are not used. So we can judge how good the algorithm is by seeing if the classes are separated.

Example 7: Displaying the structure of a set of documents using Latent Semantic Analysis (a form of PCA)



学习任务类型

Types of learning task

- Supervised learning
 - infer a function from labeled training data.
- Unsupervised learning
 - try to find hidden structure in unlabeled training data
 - clustering
- Reinforcement learning
 - To learn a policy of taking actions in a dynamic environment and acquire rewards

学习任务类型

Types of learning task

- Supervised learning
 - infer a function from labeled training data.
- Unsupervised learning
 - try to find hidden structure in unlabeled training data
 - clustering
- Reinforcement learning
 - To learn a policy of taking actions in a dynamic environment and acquire rewards

学习任务类型

Types of learning task

- Supervised learning
 - infer a function from labeled training data.



学习任务类型

Types of learning task

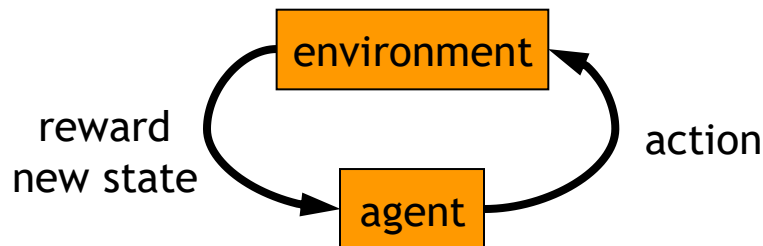
- Supervised learning
 - infer a function from labeled training data.
- Unsupervised learning
 - try to find hidden structure in unlabeled training data
 - clustering



学习任务类型

Types of learning task

- Supervised learning
 - infer a function from labeled training data.
- Unsupervised learning
 - try to find hidden structure in unlabeled training data
 - clustering
- Reinforcement learning
 - To learn a policy of taking actions in a dynamic environment and acquire rewards



学习任务类型

Types of learning task

- Supervised learning
 - infer a function from labeled training data.
- Unsupervised learning
 - try to find hidden structure in unlabeled training data
 - clustering
- Reinforcement learning
 - To learn a policy of taking actions in a dynamic environment and acquire rewards

学习任务类型

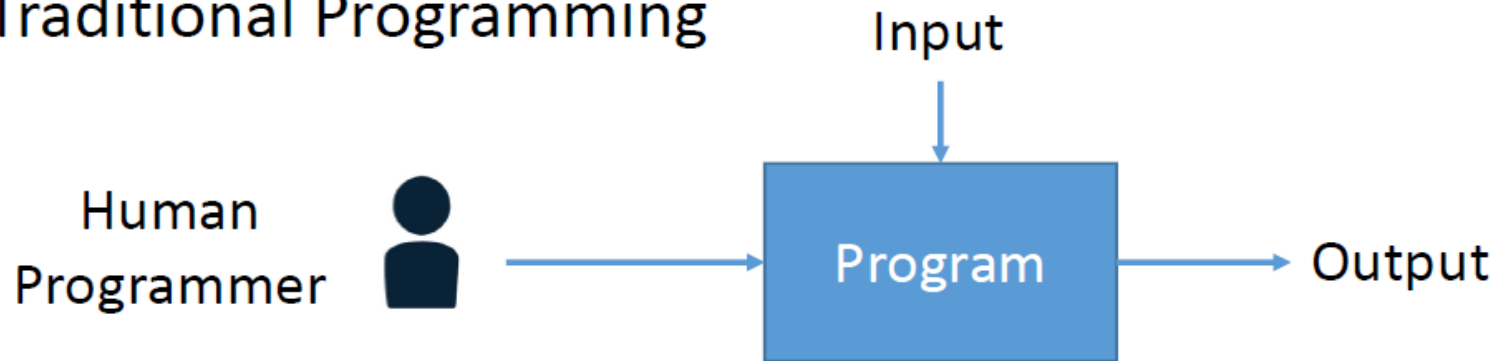
Types of learning task

		<i>Supervised Learning</i>	<i>Unsupervised Learning</i>
<i>Discrete</i>	<i>Continuous</i>	classification or categorization	clustering
		regression	dimensionality reduction

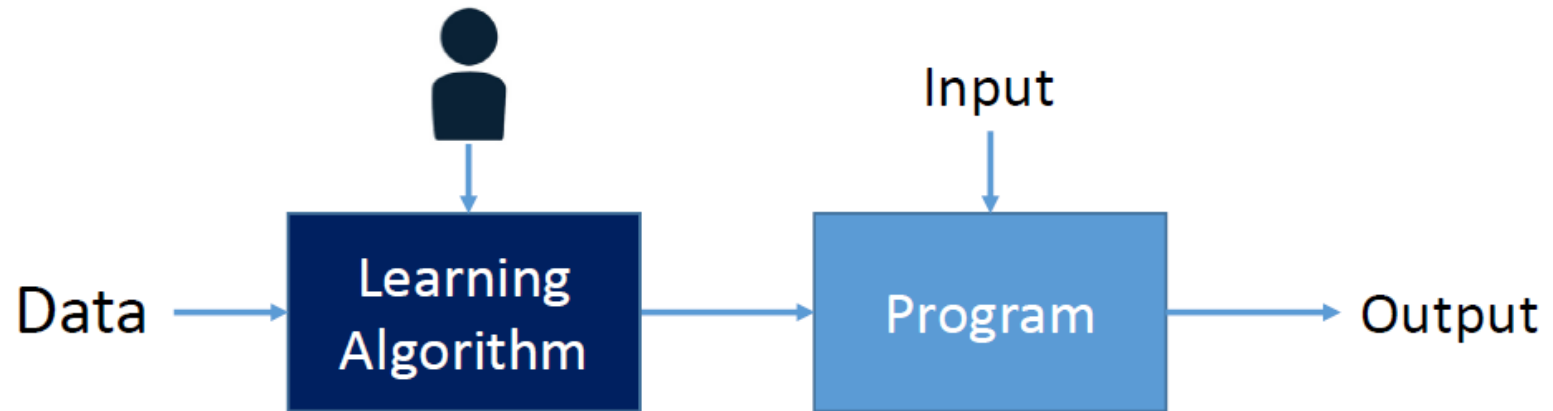
机器学习和传统编程的不同

Traditional Programming vs. Machine Learning

- Traditional Programming



- Machine Learning



机器学习发展历史

History of Machine Learning

- 1950s
 - Samuel's checker player
 - Selfridge's Pandemonium
- 1960s:
 - Neural networks: Perceptron
 - Pattern recognition
 - Learning in the limit theory
 - Minsky and Papert prove limitations of Perceptron
- 1970s:
 - Symbolic concept induction
 - Winston's arch learner
 - Expert systems and the knowledge acquisition bottleneck
 - Quinlan's ID3
 - Mathematical discovery with AM

机器学习发展历史

History of Machine Learning

- 1980s:

- Advanced decision tree and rule learning
- Explanation-based Learning (EBL)
- Learning and planning and problem solving
- Utility problem
- Analogy
- Cognitive architectures
- Resurgence of neural networks (connectionism, backpropagation)
- Valiant's PAC Learning Theory
- Focus on experimental methodology

- 1990s

- Data mining
- Adaptive software agents and web applications
- Text learning
- Reinforcement learning (RL)
- Inductive Logic Programming (ILP)
- Ensembles: Bagging, Boosting, and Stacking
- Bayes Net learning
- Support vector machines
- Kernel methods

机器学习发展历史

History of Machine Learning

- 2000s
 - Graphical models
 - Variational inference
 - Statistical relational learning
 - Transfer learning
 - Sequence labeling
 - Collective classification and structured outputs
 - Computer Systems Applications
 - Compilers
 - Debugging
 - Graphics
 - Security (intrusion, virus, and worm detection)
 - Email management
 - Personalized assistants that learn
 - Learning in robotics and vision
 - ...

机器学习发展历史

History of Machine Learning

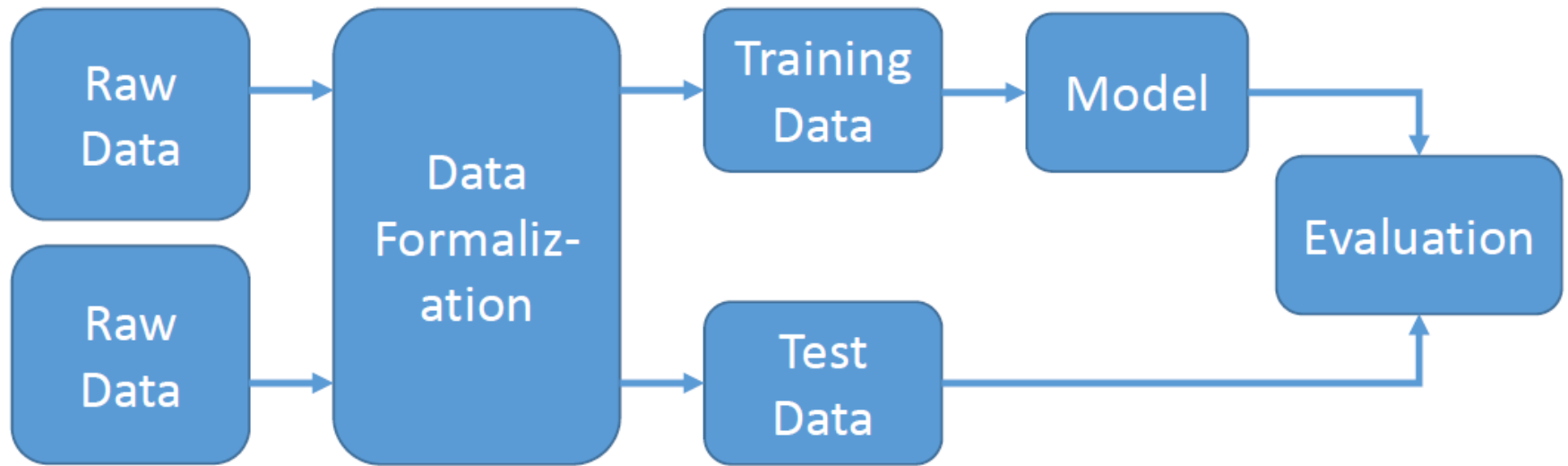
- 2010s...
 - Deep learning
 - Learning from big data
 - Learning with GPUs or HPC
 - Multi-task & lifelong learning
 - Deep reinforcement learning
 - Large Language Model

Massive applications to vision, speech, text, networks, behavior etc.

- ...

机器学习的一般过程

Machine Learning Process



- Basic assumption: there exist the same patterns across training and test data

图像分类问题举例

Image Classification Example

- Apply a prediction function to a feature representation of the image to get the desired output:

$$f(\text{apple image}) = \text{"apple"}$$

$$f(\text{tomato image}) = \text{"tomato"}$$

$$f(\text{cow image}) = \text{"cow"}$$

图像分类问题举例

Image Classification Example

$$y = f(\mathbf{x})$$

The diagram illustrates the relationship between the variables in the equation $y = f(\mathbf{x})$. Red arrows point from the labels below to the corresponding variables in the equation: from 'Output label' to y , from 'prediction function' to f , and from 'Input instance' to \mathbf{x} .

Output
label

prediction
function

Input
instance

- **Training:** given a *training set* of labeled examples $\{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \dots, (\mathbf{x}^{(N)}, \mathbf{y}^{(N)})\}$, estimate the prediction function \mathbf{f} by minimizing the prediction error on the training set
- **Testing:** apply \mathbf{f} to a never before seen test example \mathbf{x} and output the predicted value $\mathbf{y} = \mathbf{f}(\mathbf{x})$

图像分类问题举例

Image Classification Example

Training

Training
Images



Training
Labels

Image
Features

Training

Learned
model

Testing



Test Image

Image
Features

Learned
model

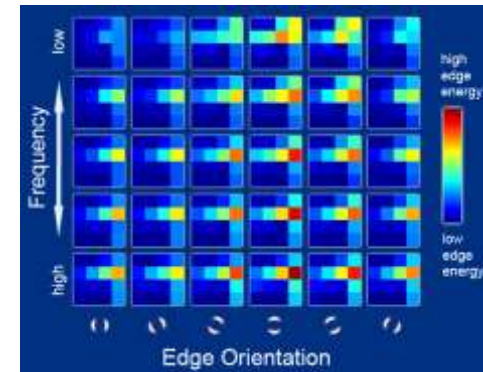
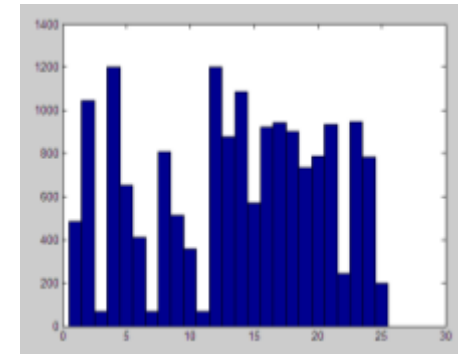
Prediction

图像分类问题举例

Image Classification Example

Features

- Raw pixels
- Histogram
- GIST descriptors
- ...



监督学习

Supervised Learning

- Given the training dataset of (data,label) pairs,

$$D = \{(x^{(i)}, y^{(i)})\}_{i=1,2,\dots,N}$$

let the machine learn a function from data to label

$$y^{(i)} \approx f_{\theta}(x^{(i)})$$

- Function set $\{f_{\theta}(x^{(i)})\}$ is called hypothesis space
- Learning is referred to as updating the parameter θ to make the prediction closed to the corresponding label

监督学习

Supervised Learning

- Given the training dataset of (data,label) pairs,

$$D = \{(x^{(i)}, y^{(i)})\}_{i=1,2,\dots,N}$$

$x^{(i)}$ = input data(features) of i^{th} training example
 $y^{(i)}$ = output data(label) of i^{th} training example

let the machine learn a function from data to label

$$y^{(i)} \approx f_{\theta}(x^{(i)})$$

- Function set $\{f_{\theta}(x^{(i)})\}$ is called hypothesis space
- Learning is referred to as updating the parameter θ to make the prediction closed to the corresponding label

监督学习

Supervised Learning

- Given the training dataset of (data,label) pairs,

$$D = \{(x^{(i)}, y^{(i)})\}_{i=1,2,\dots,N}$$

let the machine learn a function from data to label

$$y^{(i)} \approx f_{\theta}(x^{(i)})$$

- Function set $\{f_{\theta}(x^{(i)})\}$ is called hypothesis space
- Learning is referred to as updating the parameter θ to make the prediction closed to the corresponding label

1. What is the learning objective?
2. How to update the parameters?

问题1：学习目标

What is the Learning Objective?

- Make the prediction closed to the corresponding label

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N L(y^{(i)}, f_{\theta}(x^{(i)}))$$

(empirical risk minimization)

Loss function $L(y^{(i)}, f_{\theta}(x^{(i)}))$ measures the error between the label and prediction for single sample.

The definition of loss function depends on the data and task

损失函数

Loss function

□ 0-1 loss function

$$L(Y, f(X)) = \begin{cases} 1, & Y \neq f(X) \\ 0, & Y = f(X) \end{cases}$$

□ quadratic loss function

$$L(Y | f(X)) = \sum_N (Y - f(X))^2$$

□ absolute loss function

$$L(Y, f(X)) = |Y - f(X)|$$

□ logarithmic loss function、log-likelihood loss function

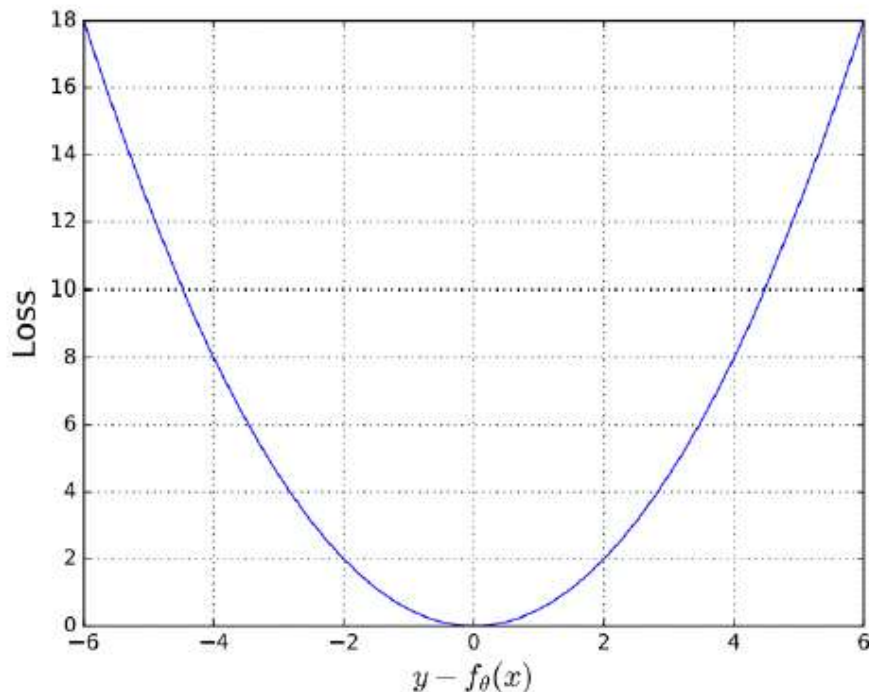
$$L(Y, P(Y | X)) = -\log P(Y | X)$$

损失函数

Loss function

Most popular loss function: squared loss

$$L(y^{(i)}, f_{\theta}(x^{(i)})) = \frac{1}{2} (y^{(i)} - f_{\theta}(x^{(i)}))^2$$



- Penalty much more on larger distances
- Accept small distance (error)
 - Observation noise etc.
 - Generalization

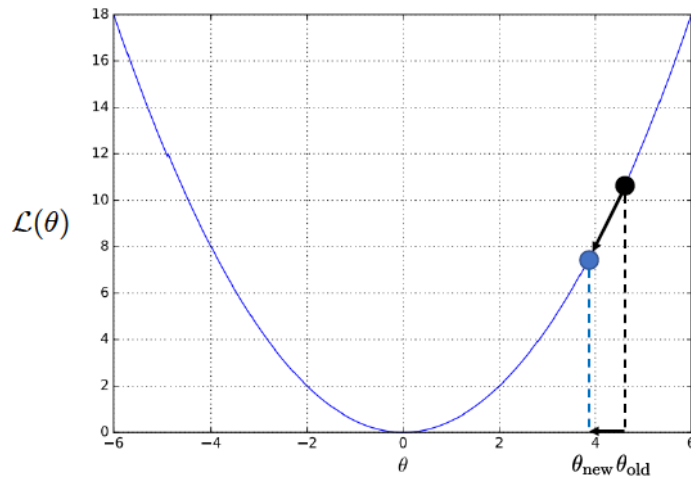
问题2：如何更新参数

How to update the parameters?

Given θ , we have code that can compute

- $J(\theta)$
- $\frac{\partial J(\theta)}{\partial \theta_i}$ (for $\theta_i \ j=0, 1, \dots, n$)

Gradient learning method:



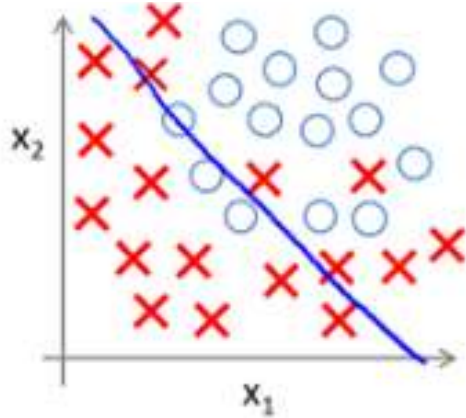
$$\theta_{\text{new}} \leftarrow \theta_{\text{old}} - \eta \frac{\partial \mathcal{L}(\theta)}{\partial \theta}$$

Other Optimization algorithms:

- Conjugate gradient method
- Quasi-Newton method
 - BFGS
 - L-BFGS
- ...

模型选择

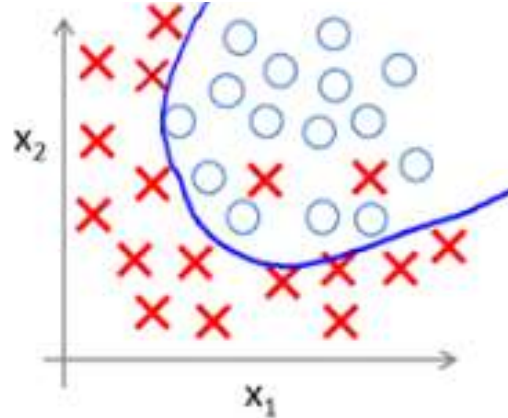
Model Selection



$$f_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

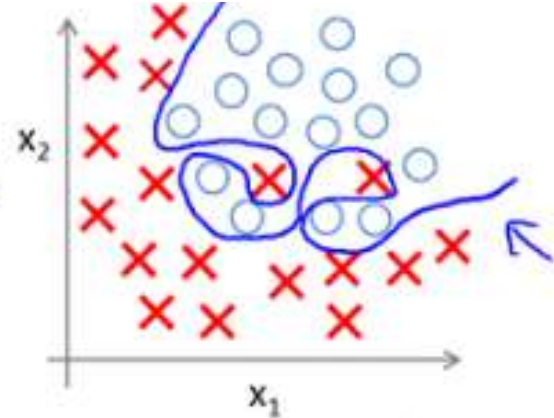
$g = \text{sigmoid function}$

Linear model: underfitting



$$f_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2)$$

Quadratic model: well fitting



$$f_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2 + \theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2 + \dots)$$

5th-order model: overfitting

- Underfitting occurs when a statistical model or machine learning algorithm cannot capture the underlying trend of the data.
- Overfitting occurs when a statistical model describes random error or noise instead of the underlying relationship

模型选择

Model Selection

Training
data



Test
data



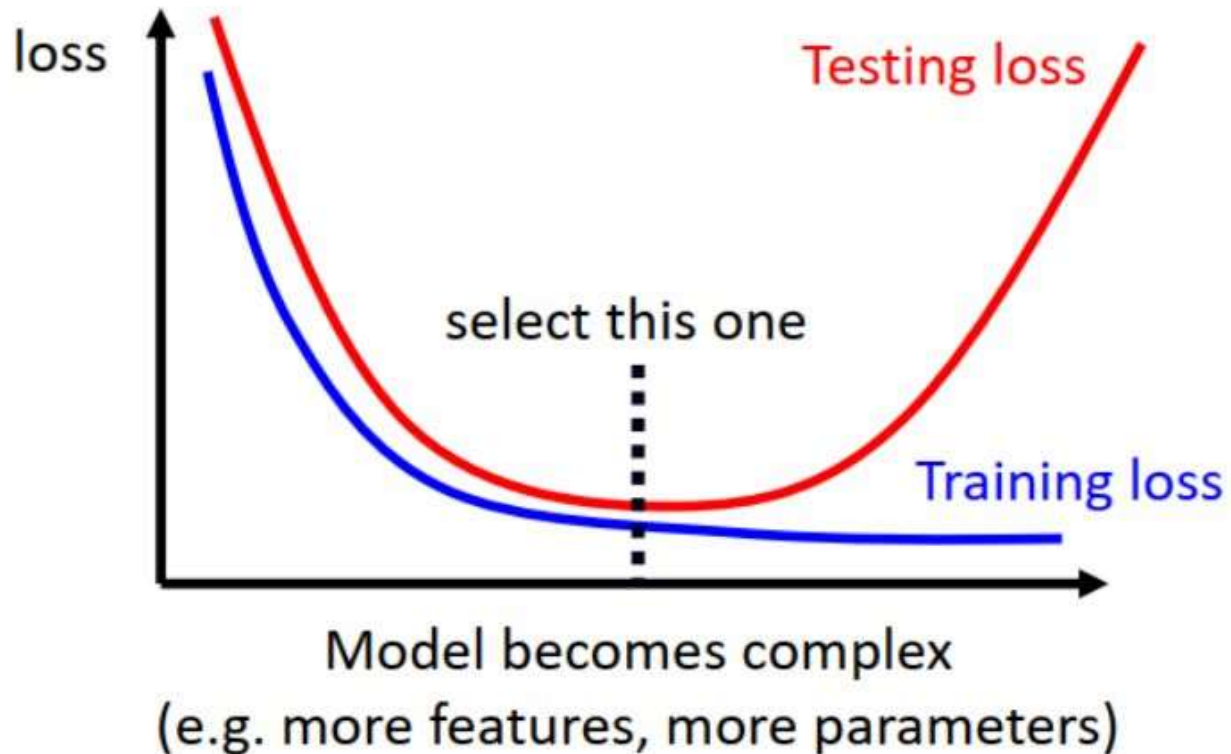
overfitting



underfitting

模型选择 Model Selection

Bias-Complexity Trade-off



奥卡姆剃刀原则

Principle of Occam's razor

Among competing hypotheses, the one with the fewest assumptions should be selected.

- Recall the function set $\{f_{\theta}(\cdot)\}$ is called hypothesis space

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N L(y^{(i)}, f_{\theta}(x^{(i)})) + \lambda \Omega(\theta)$$

Original loss

Penalty on assumptions
structural risk minimization

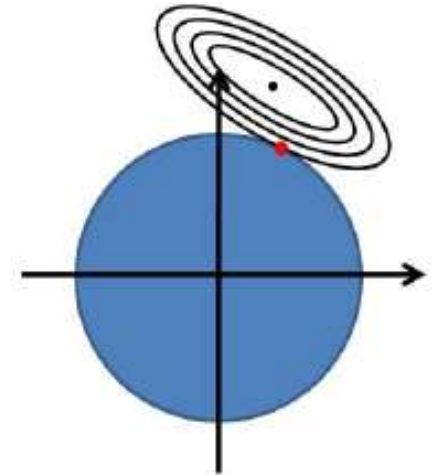
典型正则化方法

Typical Regularization

- L2-Norm (Ridge)

$$\Omega(\theta) = \|\theta\|_2^2 = \sum_{m=1}^M \theta_m^2$$

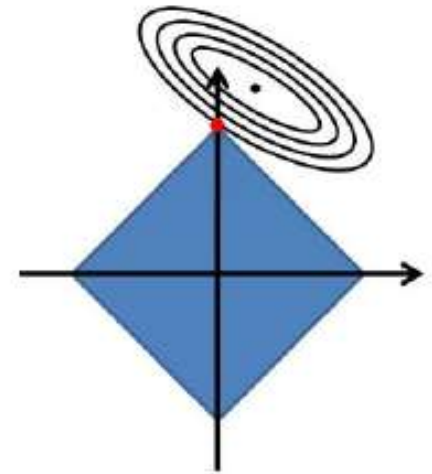
$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N L(y^{(i)}, f_{\theta}(x^{(i)})) + \lambda \|\theta\|_2^2$$



- L1-Norm (LASSO)

$$\Omega(\theta) = \|\theta\|_1 = \sum_{m=1}^M |\theta_m|$$

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N L(y^{(i)}, f_{\theta}(x^{(i)})) + \lambda \|\theta\|_1$$



典型正则化方法

Typical Regularization

L2-Norm

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N L(y^{(i)}, f_{\theta}(x^{(i)})) + \lambda \|\theta\|_2^2$$

L1-Norm

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N L(y^{(i)}, f_{\theta}(x^{(i)})) + \lambda \|\theta\|_1$$

Elastic Net:

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N L(y^{(i)}, f_{\theta}(x^{(i)})) + a \cdot (\lambda \|\theta\|_1 + (1 - \lambda) \|\theta\|_2^2)$$

损失函数&代价函数&目标函数

Loss function & Cost function & Objective function

$$L(y^{(i)}, f_{\theta}(x^{(i)}))$$

$$\frac{1}{N} \sum_{i=1}^N L(y^{(i)}, f_{\theta}(x^{(i)}))$$

$$\frac{1}{N} \sum_{i=1}^N L(y^{(i)}, f_{\theta}(x^{(i)})) + \lambda \|\theta\|_2^2$$

$$\frac{1}{N} \sum_{i=1}^N L(y^{(i)}, f_{\theta}(x^{(i)})) + \lambda \|\theta\|_1$$

模型选择

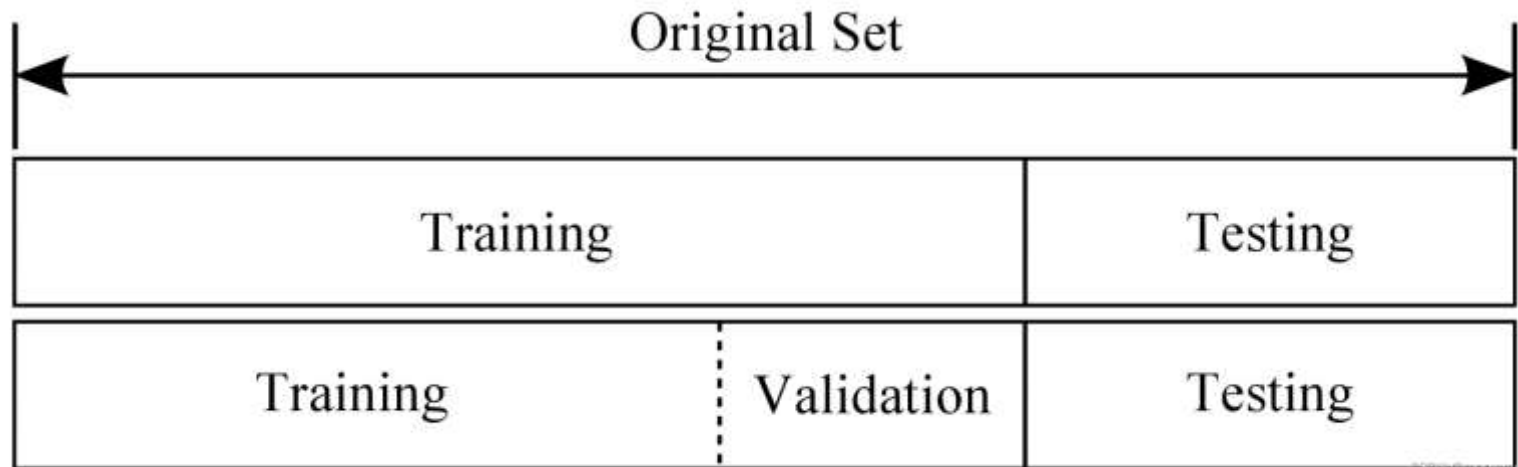
Model Selection

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N \mathcal{L}(y_i, f_{\theta}(x_i)) + \lambda \|\theta\|_2^2$$

- An ML solution has model parameters θ and optimization **hyperparameters** λ
- Hyperparameters
 - Define higher level concepts about the model such as complexity, or capacity to learn.
 - **Cannot be learned directly from the data** in the standard model training process and need to be predefined.
 - Can be decided by setting different values, training different models, and choosing the values that test better
- Model selection (or hyperparameter optimization) cares how to select the optimal hyperparameters.

留出法

hold-out for model selection



training set 60%

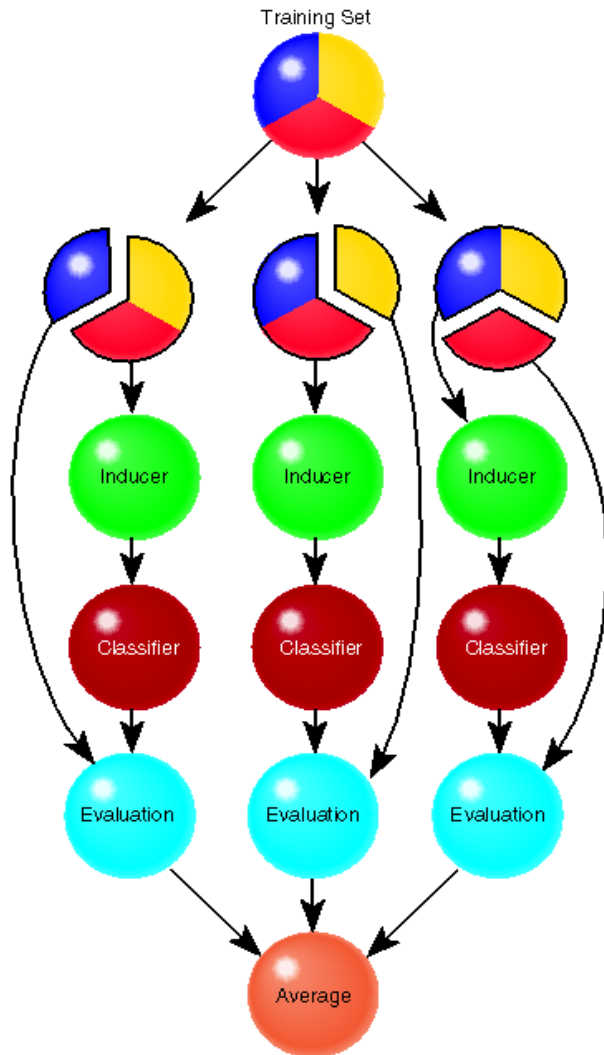
validation set 20%

testing set 20%

Split training data into
training set and validation set
for model selection

交叉验证

Cross validation for model selection



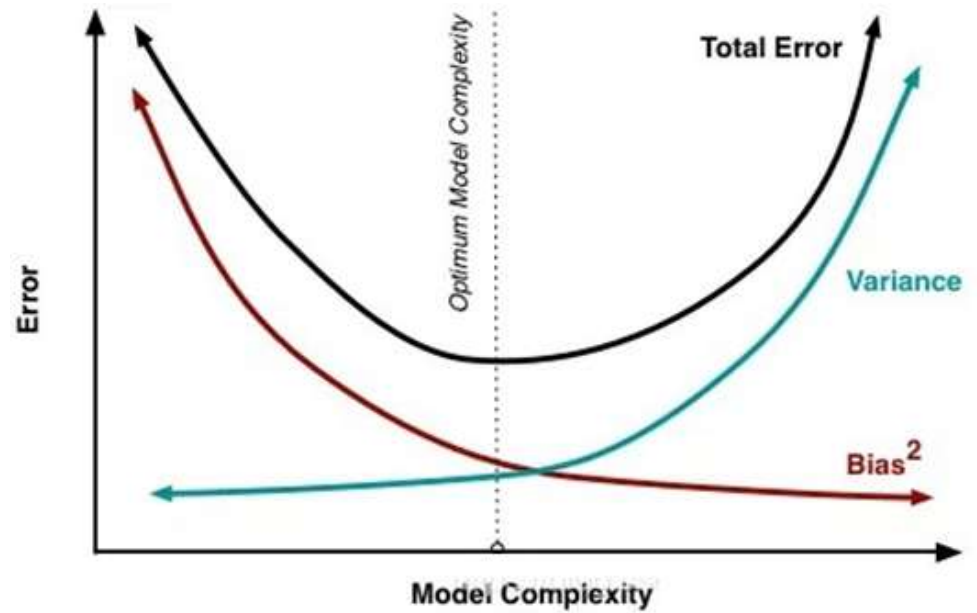
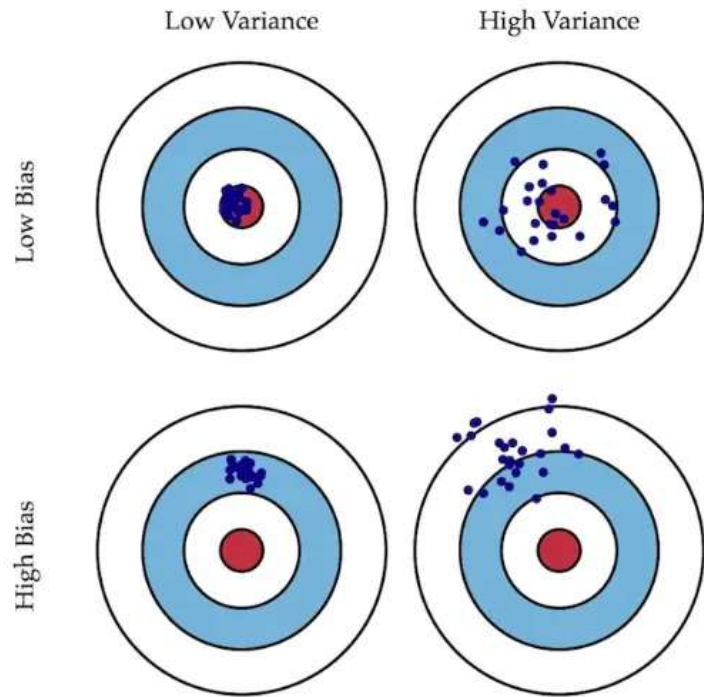
K-fold cross validation

```
CV( data S, alg L, int k )  
  Divide S into k disjoint sets { S1, S2, ..., Sk }  
  For i = 1..k do  
    Run L on S-i = S - Si  
    obtain L(S-i) = hi  
    Evaluate hi on Si  
    errSi(hi) = 1/|Si| ∑⟨x,y⟩ ∈ Si I(hi(x) ≠ y)  
  Return Average 1/k ∑i errSi(hi)
```

leave-one-out cross validation (K=|S|)

偏差和方差

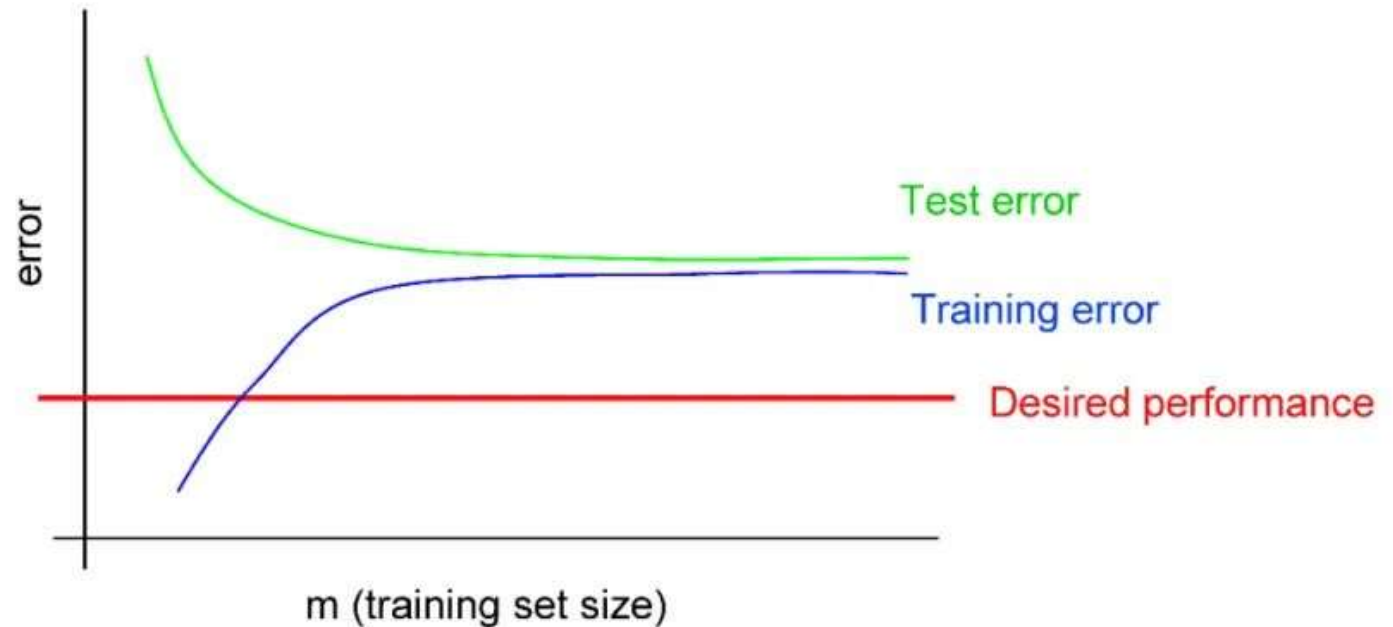
Bias & variance



偏差和方差

Bias & variance

Typical learning curve for high bias:

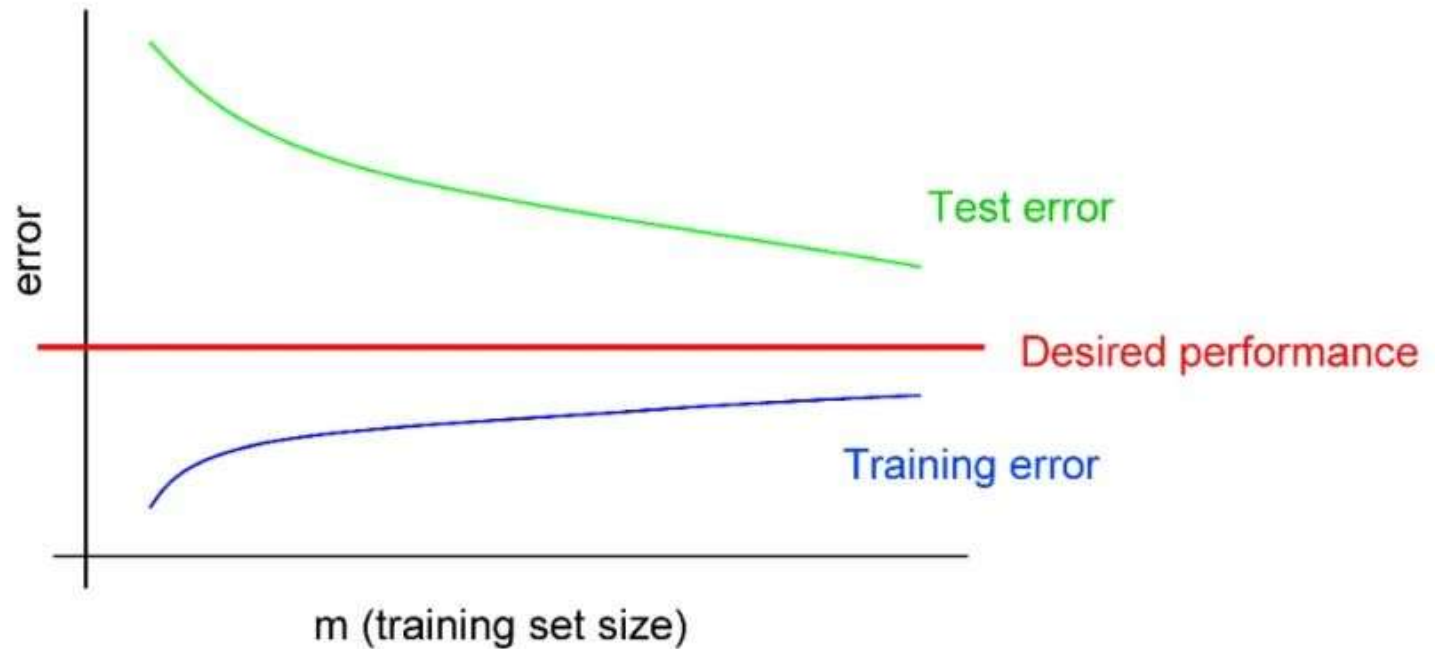


- Even training error is unacceptably high.
- Small gap between training and test error.

偏差和方差

Bias & variance

Typical learning curve for high variance:



- Test error still decreasing as m increases. Suggests larger training set will help.
- Large gap between training and test error.

如何应对欠拟合和过拟合

How to fix underfitting & overfitting

Fixes to underfitting try:

- Try a larger set of features
- Try a smaller value for λ
- Increase model complexity

Fixes to overfitting try:

- Try getting more training examples
- Try a larger value for λ
- Try a smaller set of features
- Ensemble learning

机器学习和假设空间

machine learning & Hypothesis Space

- One way to think about a supervised learning machine is as a device that **explores a “hypothesis space”**.
- The art of supervised machine learning is in:
 - Deciding how to represent the inputs and outputs
 - Selecting a hypothesis space that is powerful enough to represent the relationship between inputs and outputs but simple enough to be searched.
- **Different learning methods** assume **different hypothesis spaces** (representation languages) and/or employ **different search techniques**.

函数表征

Various Function Representations

- Numerical functions
 - Linear regression
 - Neural networks
 - Support vector machines
- Symbolic functions
 - Decision trees
 - Rules in propositional logic
 - Rules in first-order predicate logic
- Instance-based functions
 - Nearest-neighbor
 - Case-based
- Probabilistic Graphical Models
 - Naïve Bayes
 - Bayesian networks
 - Hidden-Markov Models (HMMs)
 - Probabilistic Context Free Grammars (PCFGs)
 - Markov networks

搜索算法

Various Search Algorithms

- Gradient descent
 - Perceptron
 - Backpropagation
- Dynamic Programming
 - HMM Learning
 - PCFG Learning
- Divide and Conquer
 - Decision tree induction
 - Rule learning
- Evolutionary Computation
 - Genetic Algorithms (GAs)
 - Genetic Programming (GP)
 - Neuro-evolution

课程安排

Course Arrangement

- Course content :
 - Linear models (regression)
 - Linear models (classification);
 - Decision tree ;
 - Artificial neural network;
 - Deep learning;
 - SVM;
 - Bayesian Learning
 - Ensemble learning;
 - Clustering
 - Dimension reduction& feature selection
 - Reinforcement learning
- lab assignments : regression; classification; clustering
- Final grade: usual performance15%+labs 45% + exam 40%

Textbook

- 周志华. “机器学习”. 清华大学出版社, 2016
- 李航, 统计学习方法, 清华大学出版社, 2019
- TOM M MICHELLE. Machine Learning[M]. New York: McGraw-Hill Companies, Inc, 1997.
- Andrew Ng. Machine Learning[EB/OL]. Stanford University, 2014. <https://www.coursera.org/course/ml>



吴恩达，斯坦福大学副教授，前“百度大脑”的负责人与百度首席科学家。



李航，现任字节跳动科技有限公司人工智能实验室总监，北京大学、南京大学客座教授，IEEE 会士，ACM 杰出科学家，CCF 高级员。
代表作：《统计学习方法》



周志华，南京大学计算机科学与技术系主任、人工智能学院院长。
代表作：《机器学习》（西瓜书）

Resources: Datasets

- 通用性分析：
- UCI Repository:
<http://www.ics.uci.edu/~mlearn/MLRepository.html>
- UCI KDD Archive:
<http://kdd.ics.uci.edu/summary.data.application.html>
- Statlib: <http://lib.stat.cmu.edu/>
- Delve: <http://www.cs.utoronto.ca/~delve/>
- 探索性分析：Kaggle数据集：<https://www.kaggle.com/datasets>
- 深度学习：
- [Deeplearning.net](http://deeplearning.net) - 用于对深度学习算法进行基准测试的最新数据集列表
- DeepLearning4J.org - 用于深度学习研究的高质量数据集的最新列表

Resources: Journals

- Journal of Machine Learning Research
www.jmlr.org
- Machine Learning
- Neural Computation
- Neural Networks
- IEEE Transactions on Neural Networks
- IEEE Transactions on Pattern Analysis and Machine Intelligence
- Annals of Statistics
- Journal of the American Statistical Association
- ...

Resources: Conferences

- International Conference on Machine Learning (ICML)
- Neural Information Processing Systems (NIPS)
- Computational Learning Theory (COLT)
- European Conference on Machine Learning (ECML)
- Asian Conference on Machine Learning (ECML)
- IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
- 中国机器学习大会(CCML)
- 机器学习及其应用 (MLA)

问题

- ChatGPT是一种基于Transformer架构的预训练语言模型，它的训练过程需要大量的高质量文本数据。在训练阶段，模型的主要任务是通过学习大量文本数据，从而掌握语言生成和理解的能力。
- 在预训练阶段，ChatGPT使用大量的未标注数据进行**自监督学习**，例如维基百科、新闻网站、社交媒体等。其中，主要使用了掩码建模和重建技术。掩码技术的核心思想是将模型中的一部分词语随机遮挡，然后让模型根据上下文进行预测被掩码的词语。这个任务有助于模型学习到词语之间的依赖关系和上下文信息；在上下文重建任务中，模型需要根据输入的上下文来预测下一个对话句子。这个任务有助于模型学习到上下文的语义信息的关系。
- 根据以上描述，你觉的**自监督学习**是无监督学习还是由监督学习？为什么？