

Artificial Neural Networks

人工神经网络

权小军教授
中山大学计算机学院

quanjx3@mail.sysu.edu.cn

2023 年 2 月 28 日

Lecture 1: Introduction

Lecture 1.1 Course Information

Who is this course for?

- ▶ Students who are interested in deep learning
- ▶ Students who want to know how neural networks work
- ▶ Students who want to build, train, apply a neural model
- ▶ Students who want to know the latest progress of deep learning

Deep Learning are taking over!



I have worked all my life in machine learning,
and I've never seen one algorithm knock over
benchmarks like deep learning. -Andrew Ng



Deep learning is an algorithm which has no theoretical
limitations of what it can learn; the more data you
give and the more computational time you
provide; the better it is. -Geoffrey Hinton

Deep Learning are taking over!

- ▶ Deep Learning have become one of the major thrust (推力) areas recently in various pattern recognition (模式识别), prediction, and analysis problems
- ▶ In many problems they have established the state of the art (SOTA)
 - Often exceeding previous benchmarks by large margins

Applications of deep learning: speech recognition



图: Speech recognition: from voice to text

Deep Learning are taking over!

www.technewsworld.com/story/64013.html

40 maps that explain Amazon Web Services Printers | Math n Pro... Deep Learning Tutorials deep learning PHILIPS - Golden Ears Language Technology MyIDCare - Dashboard Other bookmarks

TECHNEWSWORLD

EMERGING TECH

SEARCH

Computing Internet IT Mobile Tech Reviews Security Technology Tech Blog Reader Services

Microsoft AI Beats Humans at Speech Recognition

By Richard Adhikari Oct 20, 2016 11:40 AM PT

Print Email

5 25 45 11 0 104



Image: Adobe Stock

How do you feel about Black Friday and Cyber Monday?

- They're great -- I get a lot of bargains!
- The deals are too spread out -- I'd prefer just one day.
- They're a fun way to kick off the holiday season.
- I don't like the commercialization of Thanksgiving Day.
- They're crucial for the retail industry and the economy.
- The deals typically aren't that good.

[Vote to See Results](#)

E-Commerce Times

Black Friday Shoppers Hungry for New Experiences, New Tech

Pay TV's Newest Innovation: Giving Users Control

Apple Celebrates Itself in \$300 Coffee Table Tome

AWS Enjoys Top Perch in IaaS, PaaS Markets

US Comptroller Gears Up for Blockchain and

Applications of deep learning: language translation

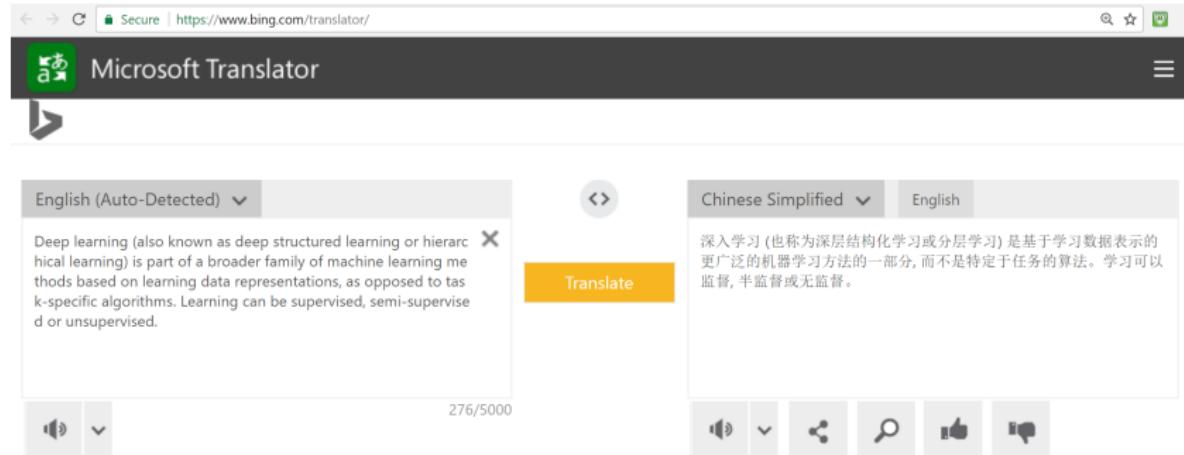
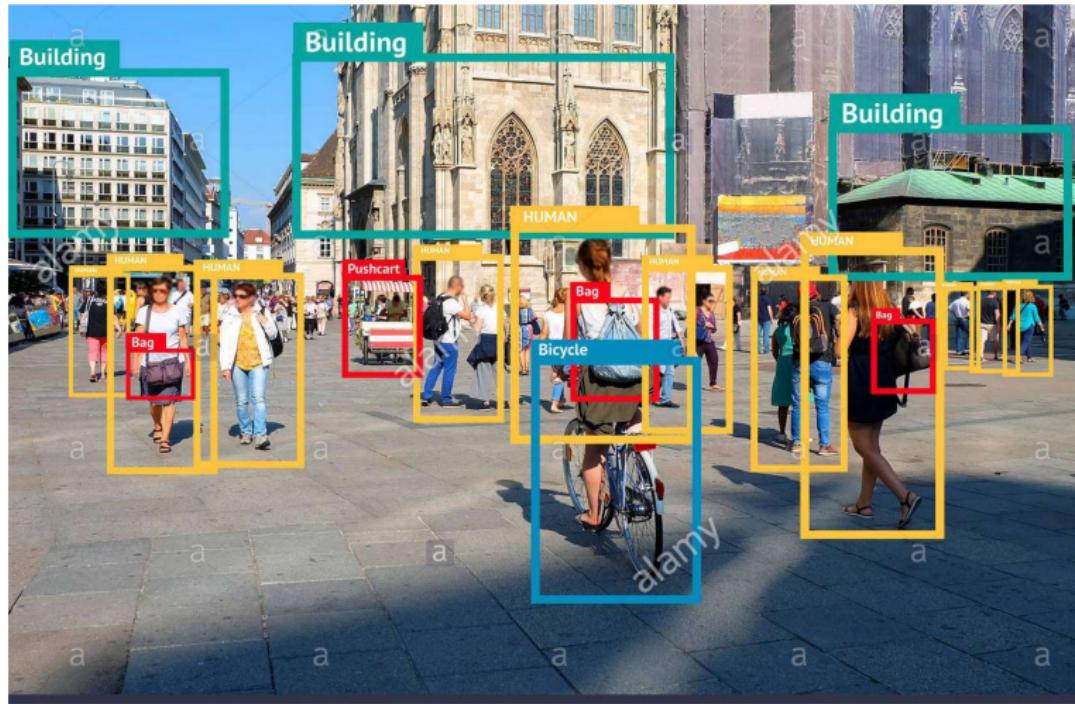
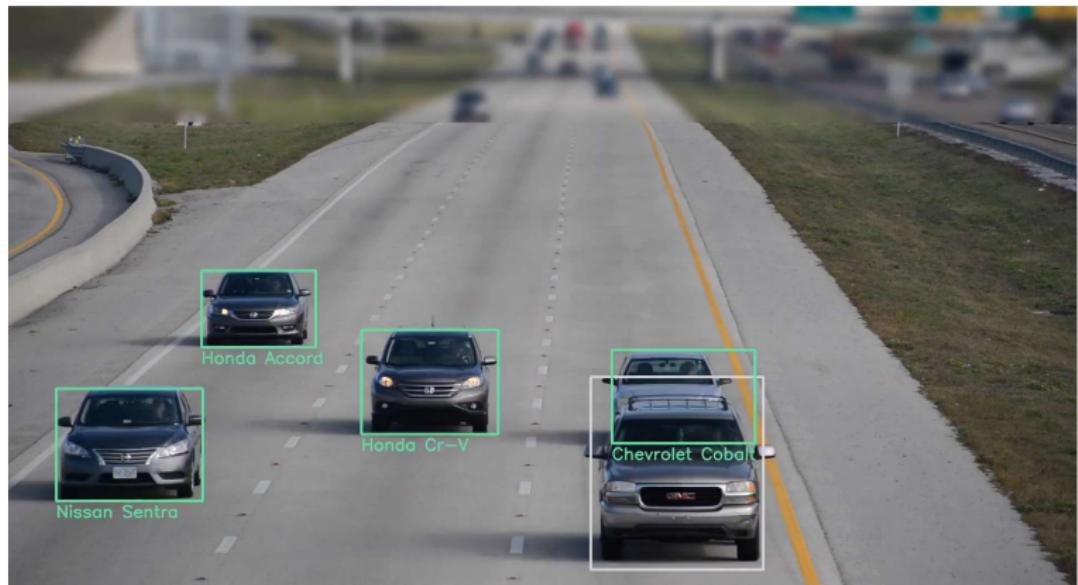


图: Machine Translation: from one language to another

Deep Learning are taking over!

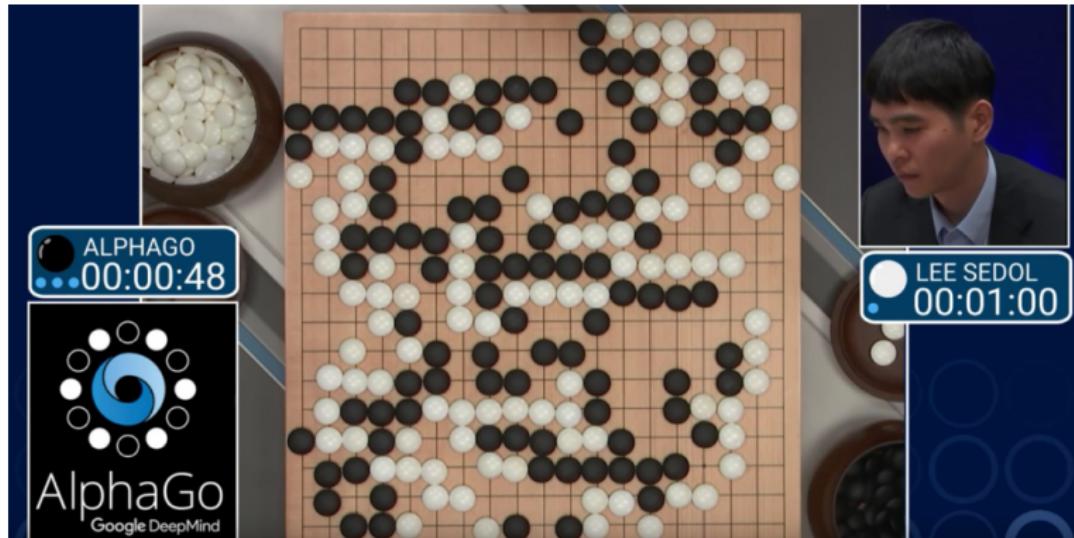


Deep Learning are taking over!



Deep Learning are taking over!

2016 年 3 月 9-15, AlphaGo 在韩国首尔对战世界围棋冠军、职业九段选手李世石



Applications of deep learning: healthcare

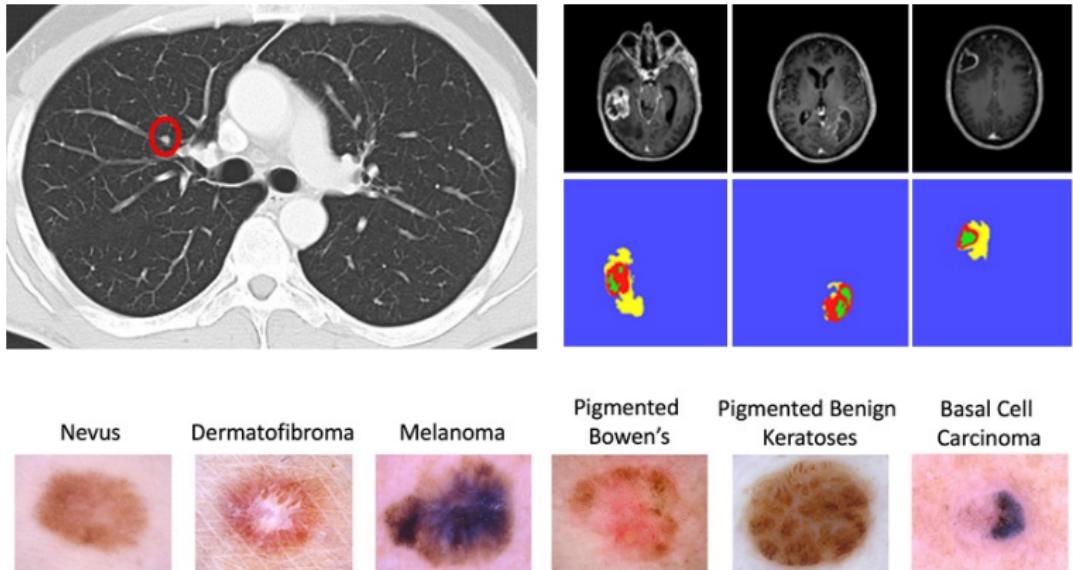


图: Medical image analysis: detection of lesions (病变), segmentation of lesion (病变) regions, diagnosis (诊断)

Applications of deep learning: pay-with-face



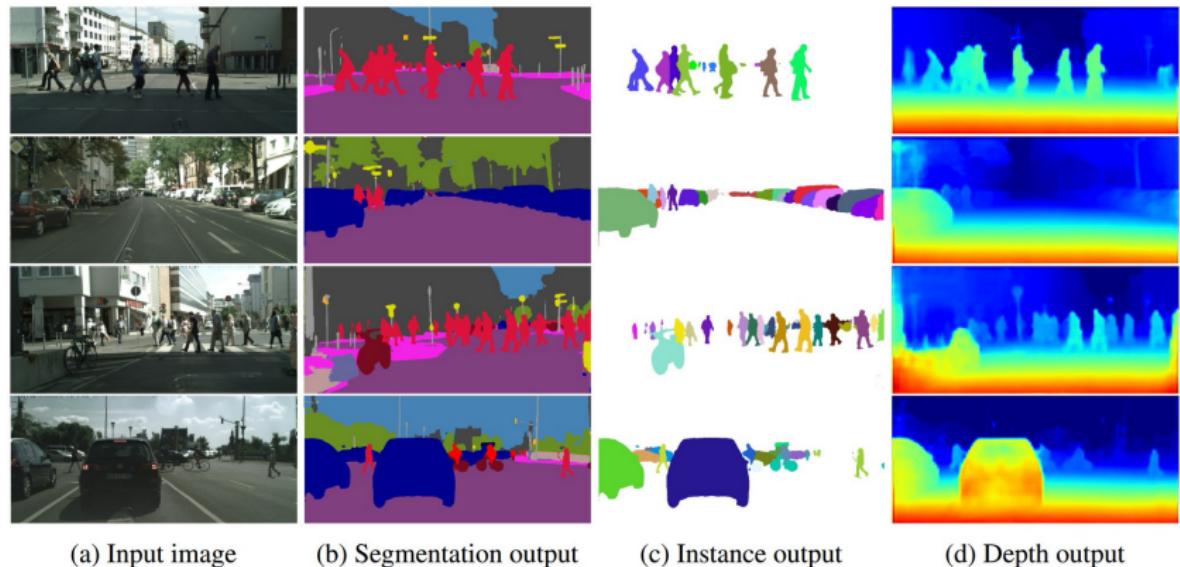
图: Face identification: classification, metric learning (距离学习), retrieval

Applications of deep learning: image recognition



图: ImageNet challenge: 1000-class image classification

Applications of deep learning: self-driving



(a) Input image

(b) Segmentation output

(c) Instance output

(d) Depth output

图: Scene understanding: semantic & instance segmentation, classification; detection of objects+humans+events, steering(转向)+acceleration(加速)

Applications of deep learning: a lot more

- ▶ Robot and chatbot (聊天机器人)
- ▶ Surveillance (监视)
- ▶ Finance
- ▶ Agriculture
- ▶ Education
- ▶ ...

Deep Learning are taking over!

Success with Deep Learning is continuing ...

What is ChatGPT?



Deep Learning are taking over!

ChatGPT: <https://chat.openai.com/chat>

- ▶ 马斯克赞美它：“好得吓人。”
- ▶ 库克赞美它：“不可思议。”
- ▶ 黄仁勋赞美它：“堪比智能手机问世。”
- ▶ 比尔·盖茨赞美它：“不亚于互联网的诞生。”

Deep Learning are taking over!

ChatGPT: <https://chat.openai.com/chat>

xi

孙悟空为什么要打死林黛玉？



这个问题似乎涉及到两个不同文化和背景的人物，孙悟空和林黛玉，他们来自不同的文学作品，因此不存在孙悟空要打死林黛玉的情节。

孙悟空是中国古代小说《西游记》中的主人公，他是一只猴子，因得道成仙而有了神通广大的能力。而林黛玉则是中国古典小说《红楼梦》中的主要人物之一，她是林家的贵族女子，文艺才华出众，但也有着悲惨的命运。

在两个作品中，并不存在孙悟空要打死林黛玉的情节，这个问题可能是一个玩笑或者误解。如果您有更多的背景信息或者疑问，请提供更多细节以便我更好地回答您的问题。

Deep Learning are taking over!

ChatGPT: <https://chat.openai.com/chat>

XI

我准备给计算机专业的大三同学讲授《人工神经网络》课程，第一节课应该讲授哪些内容？



如果您准备给计算机专业的大三同学讲授《人工神经网络》课程，那么第一节课应该包括以下内容：

1. 课程介绍和目标：简要介绍课程的内容和目标，让学生了解本课程的重点和学习收获。
2. 人工神经网络的基本概念：介绍神经网络的起源和发展历程，让学生了解什么是人工神经网络，以及神经网络在人工智能领域的重要性。
3. 神经元的基本概念：介绍神经元的结构和功能，以及激活函数的作用和类型。
4. 单层神经网络：讲解感知机模型的概念、输入、输出、权重和偏置的含义，以及使用梯度下降算法进行模型训练的基本原理。
5. 多层神经网络：介绍多层神经网络的结构和训练方法，以及前向传播和反向传播算法的基本原理。
6. 应用案例：通过实际案例介绍神经网络在图像分类、自然语言处理等领域中的应用和效果。

在第一节课中，重点是介绍人工神经网络的基本概念和原理，让学生了解神经网络的基本结构和工作原理。另外，为了让学生更好地理解和掌握所学知识，可以通过一些生动的例子和案例来帮助学生深入理解神经网络的应用和作用。

What is deep learning?

What is deep learning?

ARTIFICIAL INTELLIGENCE

Any technique that enables computers to mimic human behavior



MACHINE LEARNING

Ability to learn without explicitly being programmed



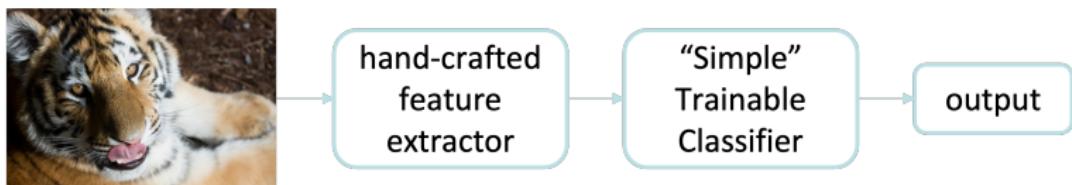
DEEP LEARNING

Extract patterns from data using neural networks



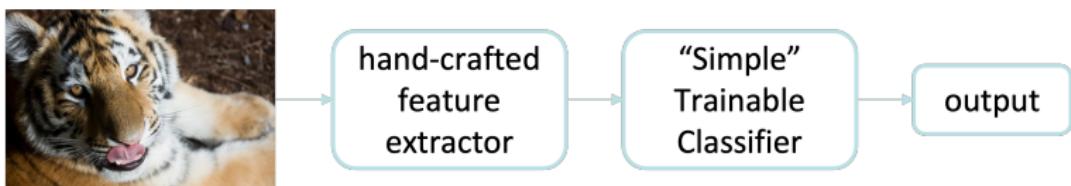
What is deep learning?

- ▶ Traditional pattern recognition (模式识别) models use hand-crafted (手工) features and simple trainable classifier.



What is deep learning?

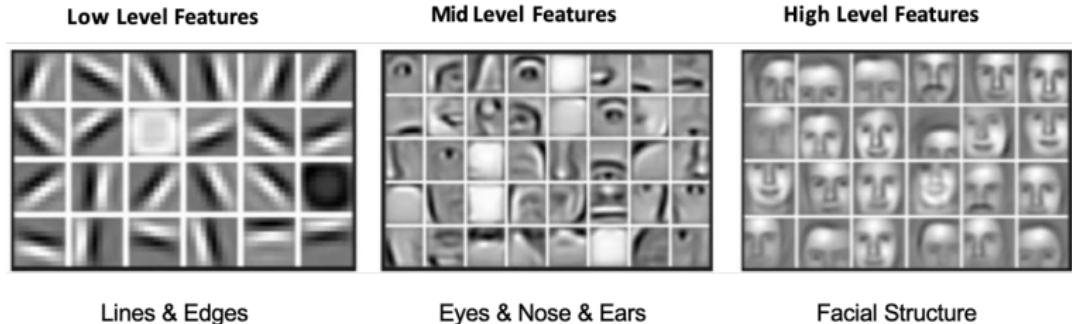
- ▶ Traditional pattern recognition (模式识别) models use hand-crafted (手工) features and simple trainable classifier.



- ▶ This approach has the following limitations:
 - ▶ It is very tedious and costly to develop hand-crafted features
 - ▶ The hand-crafted features are usually highly dependent on one application, and cannot be transferred easily to another

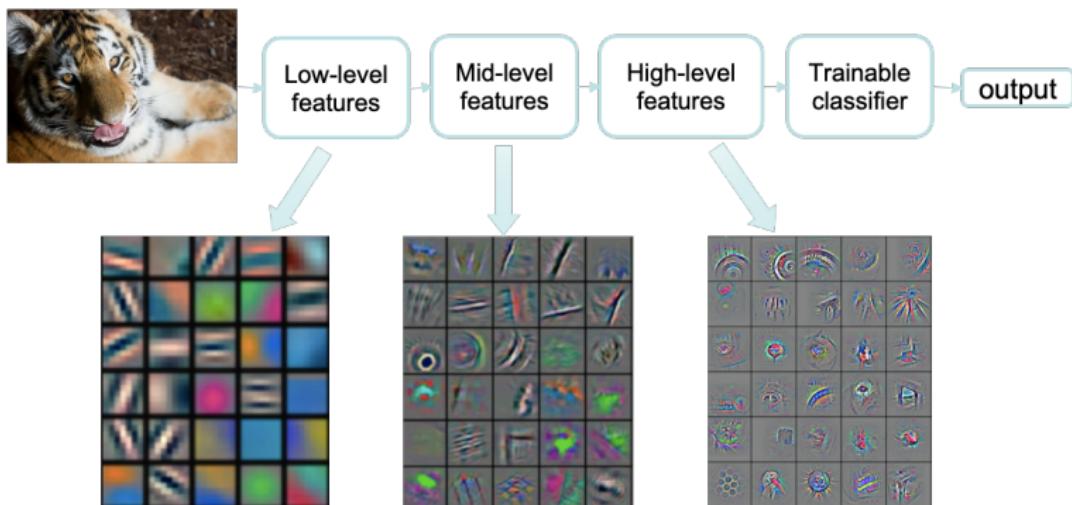
Why is deep learning?

Hand-engineered features are time consuming, brittle (脆弱的), and not scalable (可扩展) in practice. Can we learn the underlying features directly from data?



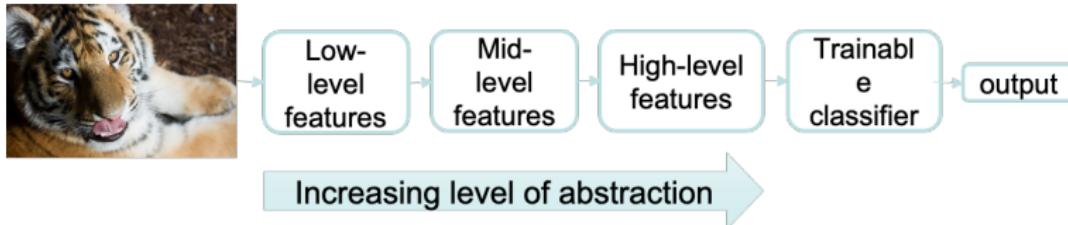
What is deep learning?

- ▶ Deep learning (a.k.a. representation learning) seeks to learn rich hierarchical representations (i.e. features) automatically through multiple stages of feature learning process.



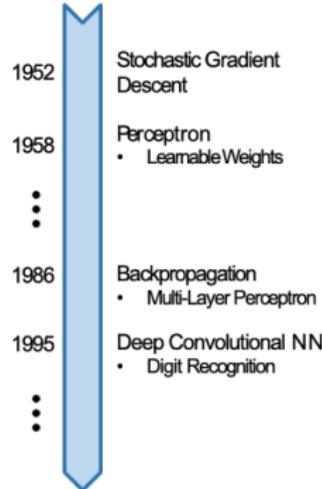
Feature visualization of convolutional net trained on ImageNet (Zeiler and Fergus, 2013)

What is deep learning?



- ▶ Hierarchy of representations with increasing level of abstraction. Each stage is a trainable nonlinear feature transform
- ▶ Image recognition
 - Pixel → edge → texton → motif → part → object
- ▶ Text
 - Character → word → phrase → clause → sentence → story

Why Now?



Neural Networks date **back** decades, so why the resurgence?

1. Big Data

- Larger Datasets
- Easier Collection & Storage

IMagenet



2. Hardware

- Graphics Processing Units (GPUs)
- Massively Parallelizable



3. Software

- Improved Techniques
- New Models
- Toolboxes



Lecture 1.2 Course Logistics

About this course

Objectives (课程目标)

- ▶ Grasp: concepts and models
- ▶ Ability/skills: design and train models
- ▶ Research: read papers/materials, and develop novel solutions for chosen projects

About this course

Objectives (课程目标)

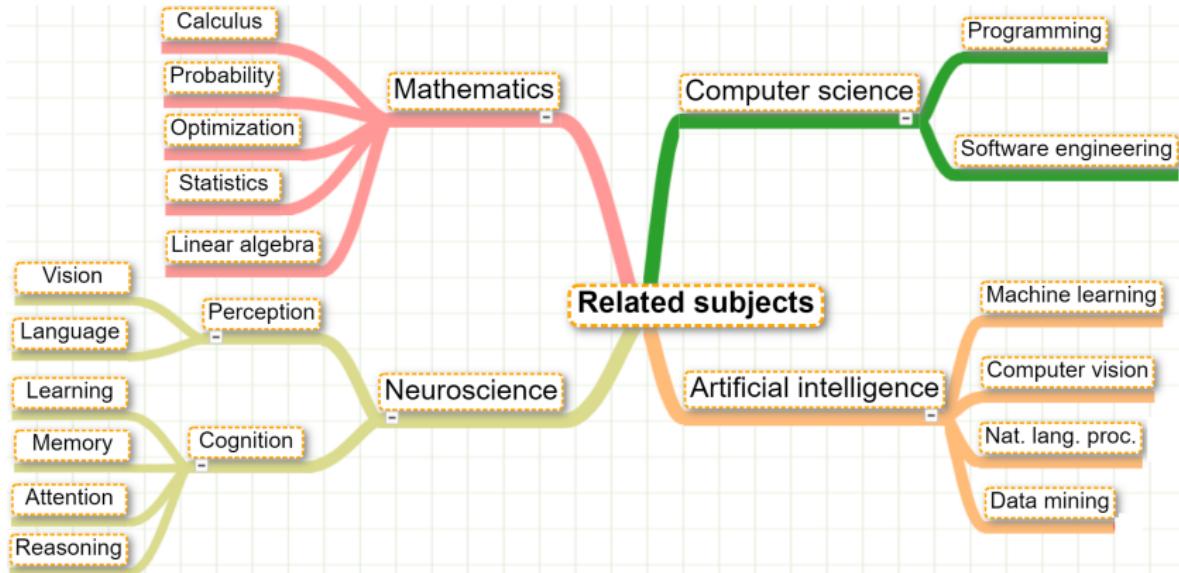
- ▶ Grasp: concepts and models
- ▶ Ability/skills: design and train models
- ▶ Research: read papers/materials, and develop novel solutions for chosen projects

Syllabus (课程大纲)

- ▶ **Lectures 1-2:** Basics of deep learning
- ▶ **Lectures 3-7:** Train and optimize neural networks
- ▶ **Lectures 8-12:** Convolutional neural networks (CNN)
- ▶ **Lectures 13-18:** Recurrent neural networks (RNN)
- ▶ **Lectures 19-25:** Advanced topics

No lab class, but we will provide lab material for you to practice!

About this course



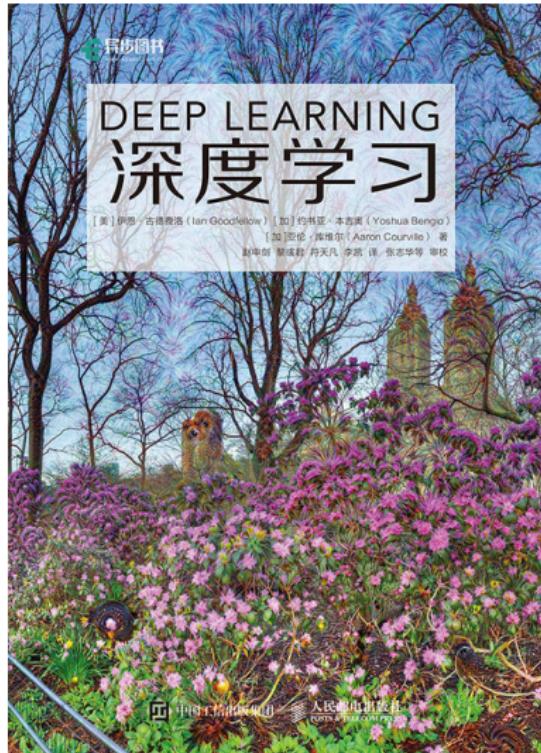
About this course (cont')

Grading

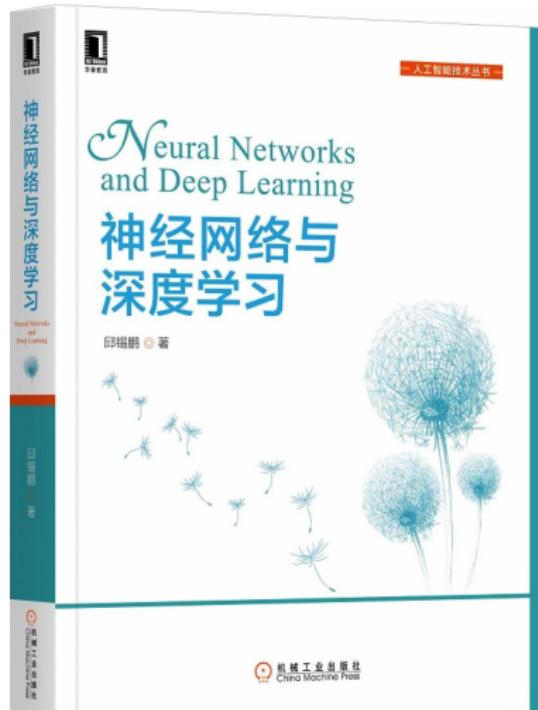
- | | |
|------------------|--|
| Assignment (40%) | ▶ Coding: implementation
▶ In-class quiz
▶ Report summary of materials outside class |
| Project (40%) | ▶ One final project
▶ Report: formal, Chinese is ok, but English is preferable
▶ Code submission |
| Other (10+10%) | ▶ Attendance; Q&A, ideas/thoughts, etc. |

Reference books

<https://www.deeplearningbook.org/>



Reference books



Wechat and lecture slides

群聊: 2023春《人工
神经网络》



(a) Wechat Group



(b) Lecture Slides

Lecture 1.3 A Machine Learning Example

A linear regression problem: an example

A simple problem

Develop a system to predict a person's weight based on his height.

A linear regression problem: an example

A simple problem

Develop a system to predict a person's weight based on his height.

To solve this problem

1. **Formulation:** suppose a person's weight y is linearly correlated with his height x , i.e., $y = wx + b$

A linear regression problem: an example

A simple problem

Develop a system to predict a person's weight based on his height.

To solve this problem

1. **Formulation:** suppose a person's weight y is linearly correlated with his height x , i.e., $y = wx + b$
2. **Data:** collect a number of persons' weight and height:
$$D = \{(x_i, y_i) | i = 1 \dots N\}$$

A linear regression problem: an example

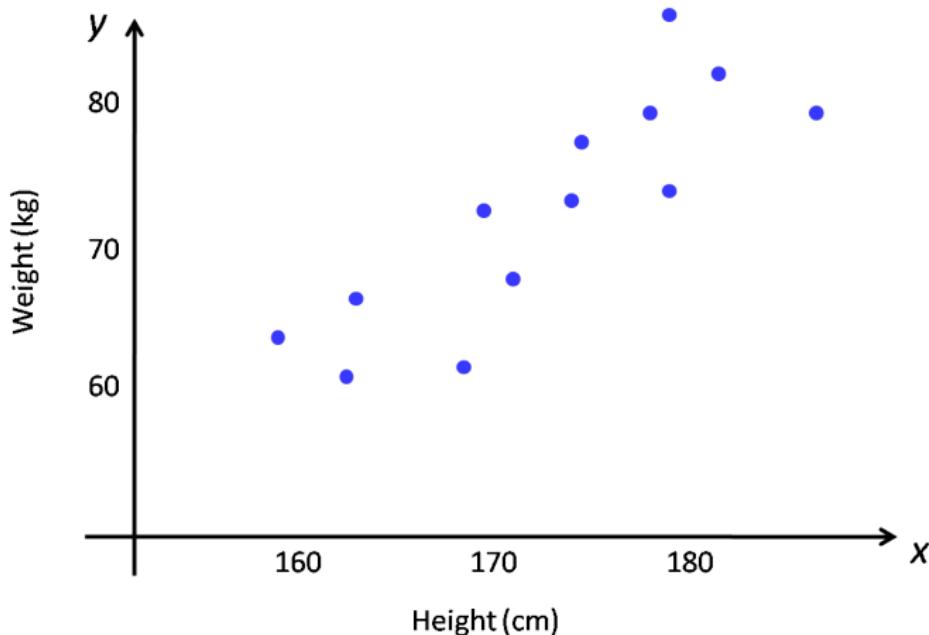
A simple problem

Develop a system to predict a person's weight based on his height.

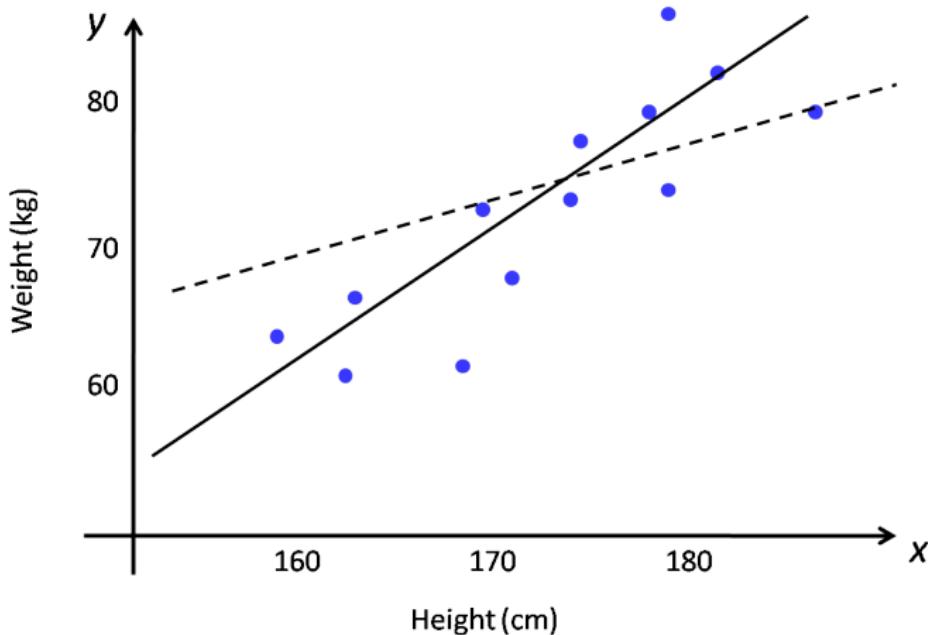
To solve this problem

1. **Formulation:** suppose a person's weight y is linearly correlated with his height x , i.e., $y = wx + b$
2. **Data:** collect a number of persons' weight and height:
 $D = \{(x_i, y_i) | i = 1 \dots N\}$
3. **Optimization:** use D to find best parameters $\theta^* = \{w^*, b^*\}$

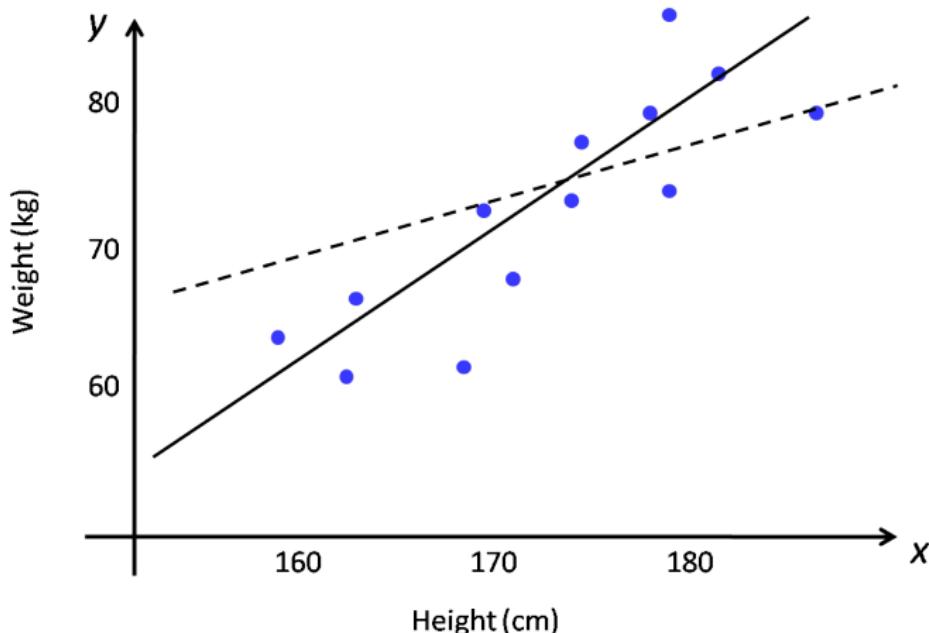
Linear regression



Linear regression (cont')

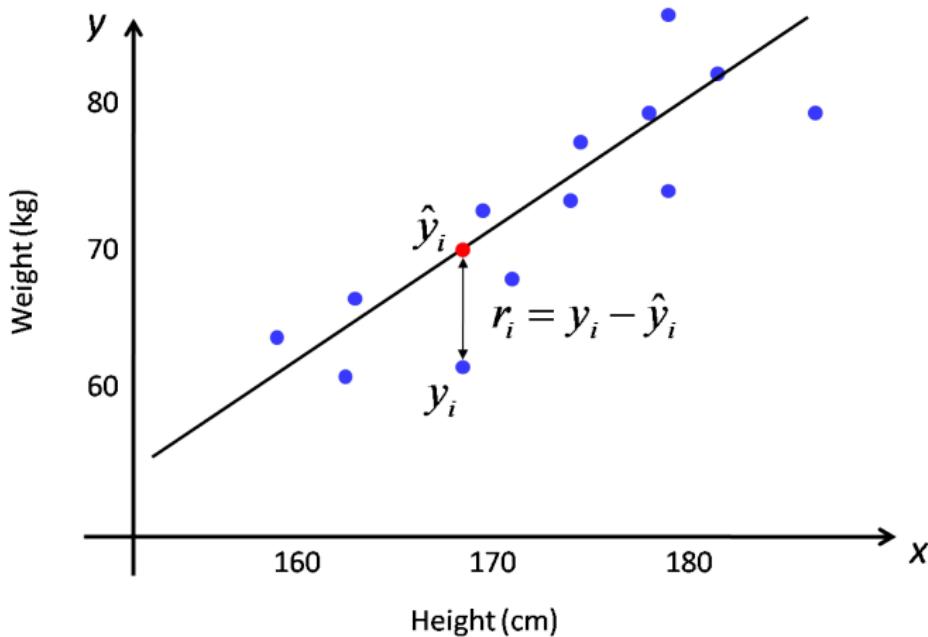


Linear regression (cont')

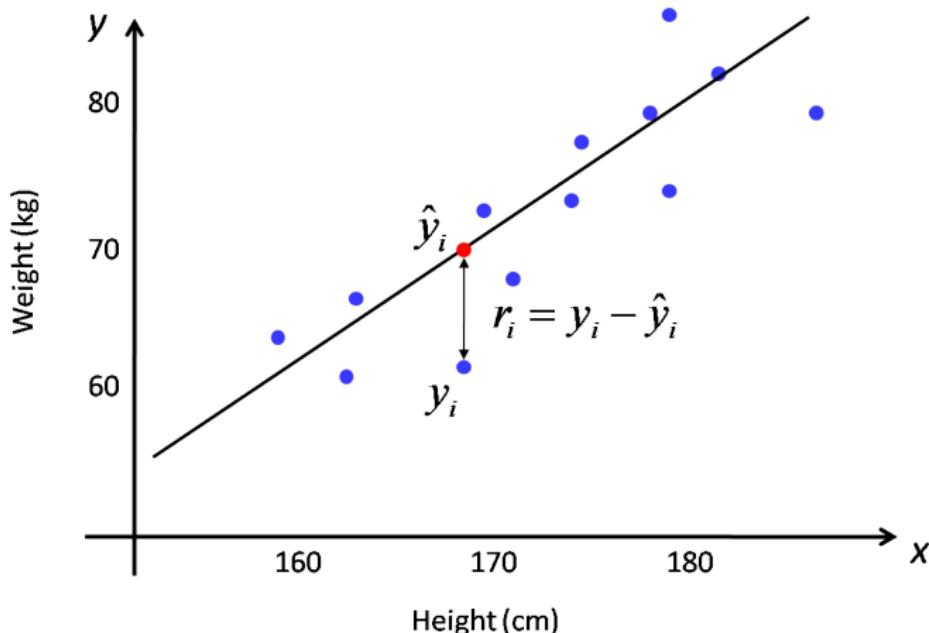


Rule: predicted line should be as close to observed points as possible.

Linear regression (cont')



Linear regression (cont')



Rule rephrased: Minimization of sum of squared difference r_i^2

Linear regression (cont')

Minimization of sum of squared difference r_i^2

$$\begin{aligned}\min_{\boldsymbol{\theta}} \quad L(\boldsymbol{\theta}) &= \frac{1}{N} \sum_{i=1}^N r_i^2 \\ &= \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \\ &= \frac{1}{N} \sum_{i=1}^N (y_i - wx_i - b)^2\end{aligned}$$

- $\boldsymbol{\theta} = \{w, b\}$ model parameters; $L(\boldsymbol{\theta})$ loss function

Linear regression (cont')

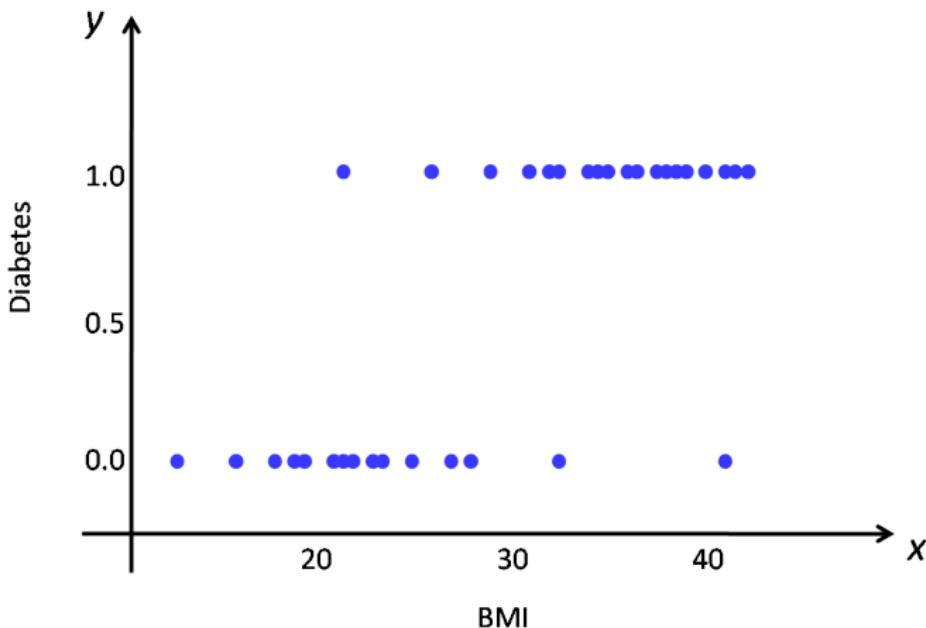
Minimization of sum of squared difference r_i^2

$$\begin{aligned}\min_{\theta} L(\theta) &= \frac{1}{N} \sum_{i=1}^N r_i^2 \\ &= \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \\ &= \frac{1}{N} \sum_{i=1}^N (y_i - wx_i - b)^2\end{aligned}$$

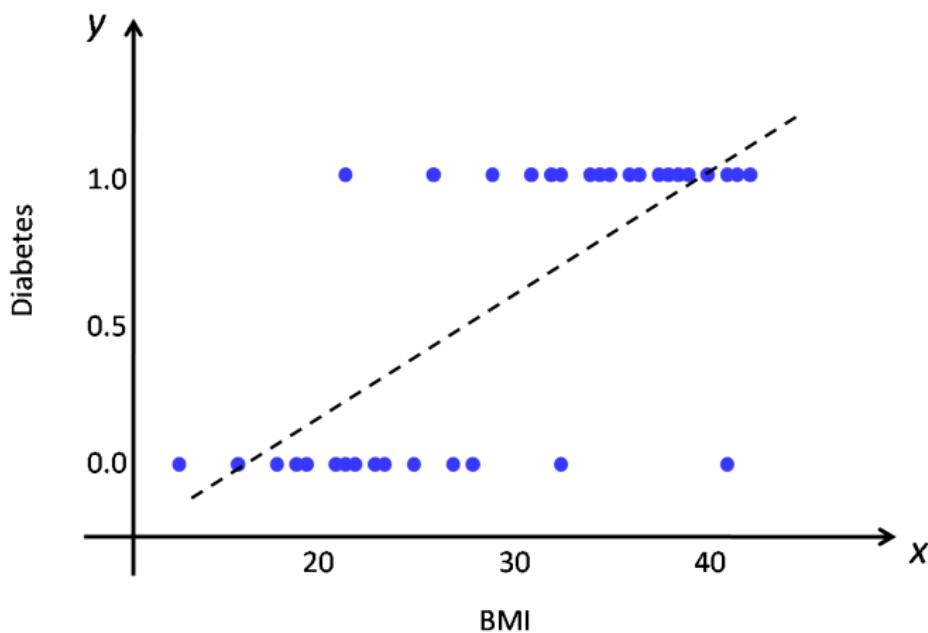
- ▶ $\theta = \{w, b\}$ model parameters; $L(\theta)$ loss function
- ▶ Solving a problem = optimization (min.) of a loss function with the help of *training data* to find best model parameters!

How to solve??

What if predicted variable is binary?

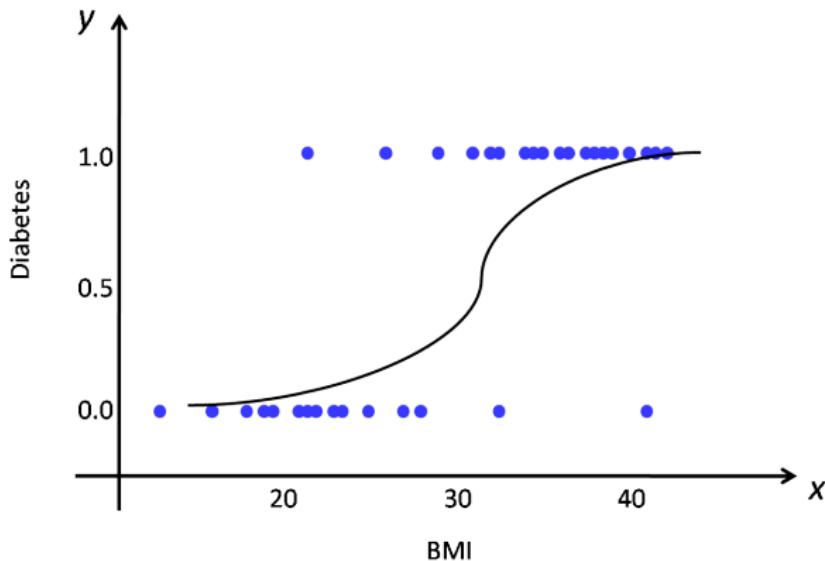


What if predicted variable is binary? (cont')



A linear model is a bad choice!

What if predicted variable is binary? (cont')



A logistic regression (逻辑回归) model is often used:

$$y = \frac{1}{1 + e^{-(wx+b)}} \quad (1)$$

Further: what if predicted variable is categorical?

Classification

- ▶ Given an input data (e.g., an image, a video clip, recorded voice sequence, a document, etc.), predict the category or class label of the data
- ▶ Number of categories: two or more

Further: what if predicted variable is categorical?

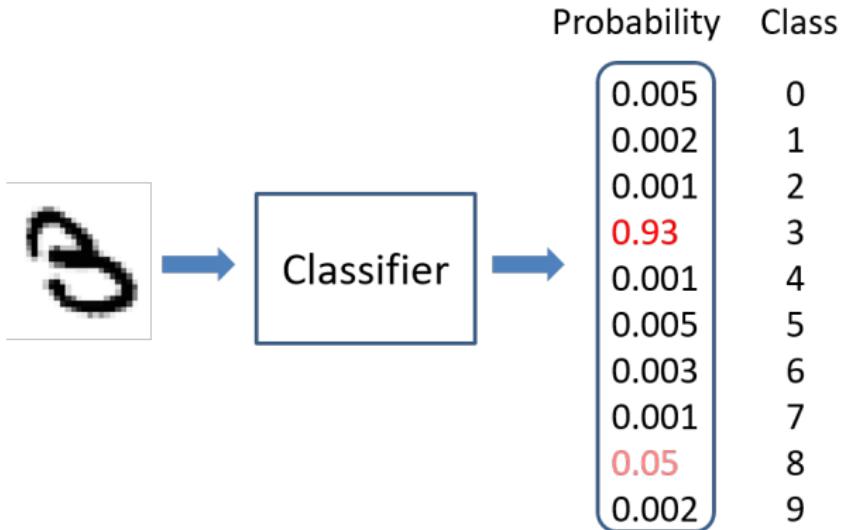
Classification

- ▶ Given an input data (e.g., an image, a video clip, recorded voice sequence, a document, etc.), predict the category or class label of the data
- ▶ Number of categories: two or more

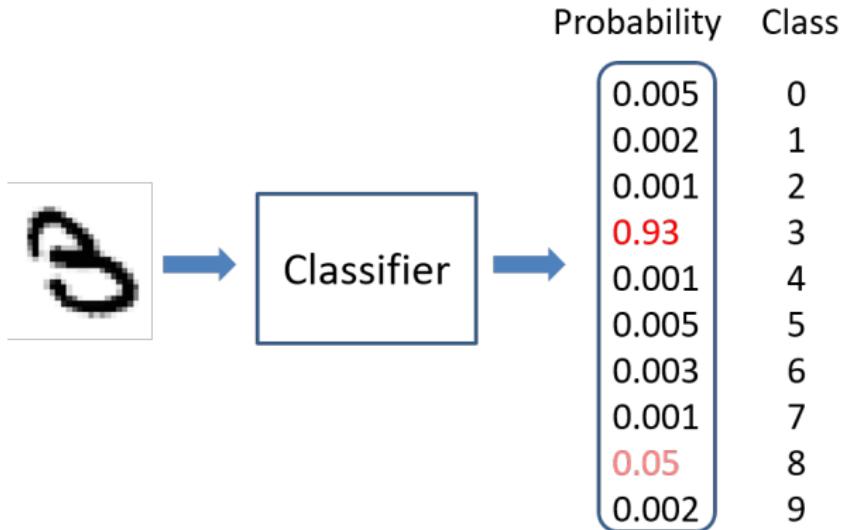
Classifier

- ▶ Given an input data, output the probability of the data belonging to each class
- ▶ So, classifier output is often a probability distribution, i.e., not a single value but a set of values

What if predicted variable is categorical? (cont')



What if predicted variable is categorical? (cont')



How to make sure the classifier output is a probability distribution?

What if predicted variable is categorical? (cont')

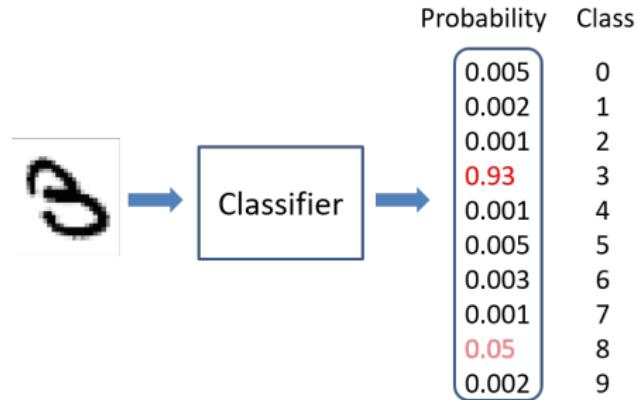
Softmax function

- ▶ Transformation of a K -dimensional vector \mathbf{z} to another K -dimensional vector $\sigma(\mathbf{z})$ in the range $[0,1]$
- ▶ $\sigma(\mathbf{z})$ is a probability distribution: sum up to 1
- ▶ Generalization of the logistic function
- ▶ Also called *normalized exponential function*

$$\begin{aligned}\sigma : \quad \mathbb{R}^K &\rightarrow [0, 1]^K \\ \sigma(\mathbf{z})_j &= \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}} \quad \text{for } j = 1, \dots, K.\end{aligned}$$

To obtain a classifier

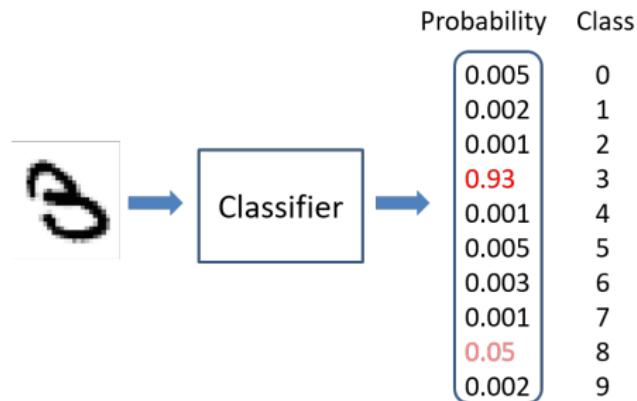
- ▶ How to design a classifier model $y = f(x; \theta)$?



- ▶ How to teach (i.e., train) the classifier?

To obtain a classifier

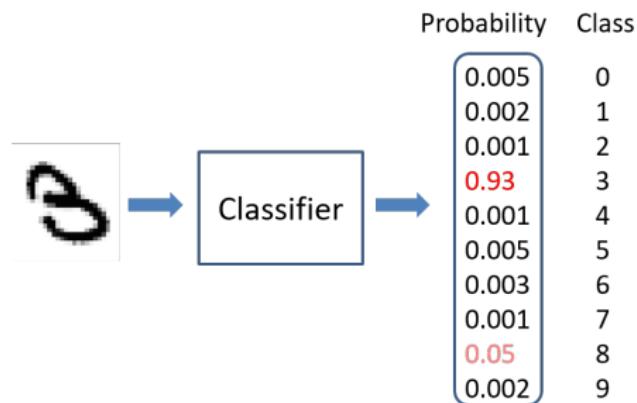
- ▶ How to design a classifier model $y = f(x; \theta)$?
 $f(x; \theta)$ could be an **artificial neural network!**



- ▶ How to teach (i.e., train) the classifier?

To obtain a classifier

- ▶ How to design a classifier model $y = f(x; \theta)$?
 $f(x; \theta)$ could be an **artificial neural network!**



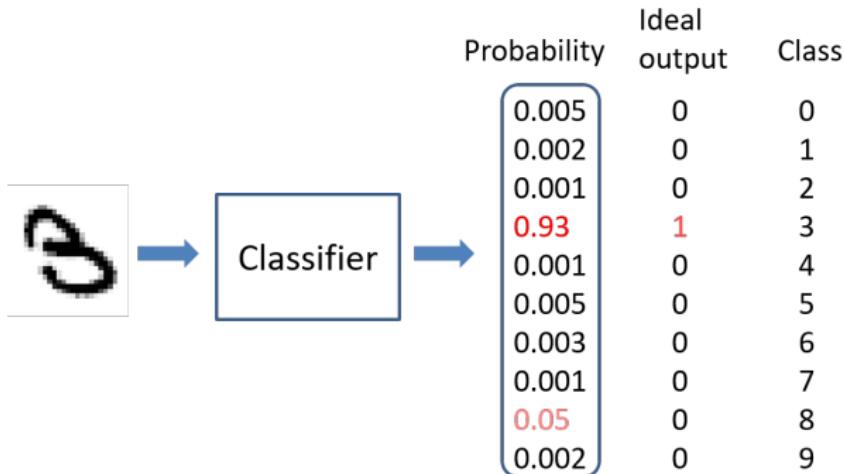
- ▶ How to teach (i.e., train) the classifier?
Design a loss function, then minimize it!

Classifier training

- ▶ Goal: find optimal parameters θ^* of a classifier $y = f(x; \theta)$
- ▶ Rule: given input x , classifier output $f(x; \theta)$ should be as close to the *ideal output* as possible.

Classifier training

- ▶ Goal: find optimal parameters θ^* of a classifier $y = f(x; \theta)$
- ▶ Rule: given input x , classifier output $f(x; \theta)$ should be as close to the *ideal output* as possible.



Cross entropy loss 交叉熵损失

- ▶ Classifier output and ideal output are represented as two probability distributions.
- ▶ How to measure difference between two distributions?

Cross entropy loss 交叉熵损失

- ▶ Classifier output and ideal output are represented as two probability distributions.
- ▶ How to measure difference between two distributions?
Cross entropy!

Cross entropy loss 交叉熵损失

- ▶ Classifier output and ideal output are represented as two probability distributions.
- ▶ How to measure difference between two distributions?
Cross entropy!
- ▶ Given input \mathbf{x}_i ,

Classifier output $\hat{\mathbf{y}}_i = \mathbf{f}(\mathbf{x}_i; \boldsymbol{\theta}) = (\hat{y}_{i1}, \hat{y}_{i2}, \dots, \hat{y}_{iK})$,

Ideal output $\mathbf{y}_i = (y_{i1}, \dots, y_{iK}) = (0, \dots, 1, \dots, 0)$,

$$\begin{aligned} l(\mathbf{y}_i, \mathbf{f}(\mathbf{x}_i; \boldsymbol{\theta})) &= \sum_{k=1}^K y_{ik} \log \frac{1}{\hat{y}_{ik}} \\ &= - \sum_{k=1}^K y_{ik} \log \hat{y}_{ik} \end{aligned}$$

Cross entropy loss to train a classifier

- ▶ Collect a set of training data $D = \{(\mathbf{x}_i, \mathbf{y}_i) | i = 1, \dots, N\}$
- ▶ Rule: find best parameters with minimum cross entropy loss.

$$\begin{aligned}\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) &= \frac{1}{N} \sum_{i=1}^N l(\mathbf{y}_i, \mathbf{f}(\mathbf{x}_i; \boldsymbol{\theta})) \\ &= -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K y_{ik} \log \hat{y}_{ik}\end{aligned}$$

Minimizing loss function: gradient descent

Gradient descent

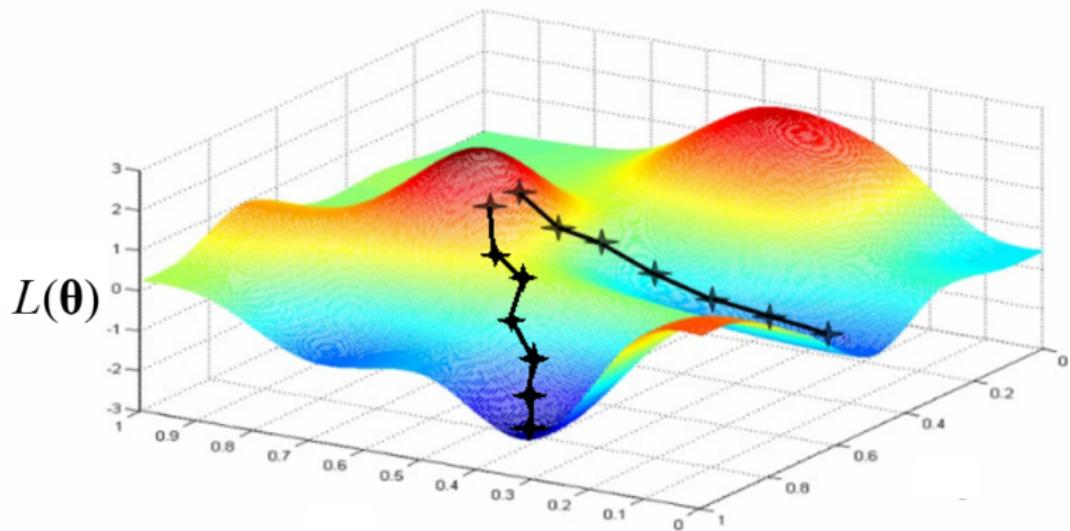
An iterative process:

1. Start from random parameters θ_0
2. For the $(t + 1)^{th}$ iteration, compute gradient $\nabla L(\theta_t)$
3. Update parameters $\theta_{t+1} = \theta_t - r \nabla L(\theta_t)$

Suppose $\theta_t = (\theta_{t,1}, \theta_{t,2}, \dots, \theta_{t,D})^T$, gradient of loss function $L(\theta)$ at θ_t is defined as $\nabla L(\theta_t) = (\frac{\partial L}{\partial \theta_{t,1}}, \frac{\partial L}{\partial \theta_{t,2}}, \dots, \frac{\partial L}{\partial \theta_{t,D}})^T$.

r is learning rate, a hyper-parameter (超参) to be tuned.

Gradient descent can help find local minimum!



Summary

- ▶ ANN is a much simplified version of human brain
- ▶ ‘Deep’ here means ‘multi-layers’
- ▶ This course: knowledge, practice, and creativity
- ▶ How to solve a problem? Formulation, data, optimization
- ▶ Gradient descent can be used to train a neural network.

Further reading: Chapter 1, in textbook “Deep learning”,
<http://www.deeplearningbook.org/>

Thank you!