Lecture 22

§1 Machine learning

1. What is machine learning (机器学习)

1° 建造能 ① automatically 从历史数据中学习，② identify patterns，③ 在无人类 intervention (干预) 下做出 logic decisions 的 system.

2° input data 有多种形式：数字、文字、点击、图象 ---.

3° Machine learning applications 可以通过 automated optimization methods 来 learn from the input data，并持续提升 outputs 的准确性.

4° Machine learning 的质量主要取决于两方面：

- input data 的质量
- The model choice itself. (每个算法有自己的 specific uses)

2. Why is machine learning important

- Machine learning is growing in importance due to increasingly enormous volumes and variety of data, the access and affordability of computational power, and the availability of high speed Internet.
- It is possible for one to rapidly and automatically develop models that can quickly and accurately analyze extraordinarily large and complex data sets.
- Many applications: cut costs, mitigate risks, and improve overall quality of life including recommending products/services, detecting cybersecurity breaches, and enabling self-driving cars.

3. How does machine learning work

1° Step 1: 选择并准备一个 training data set (训练集)

training data: 有代表性的信息，机器学习软件可以摄取并调整模型系数.

2° Step 2: 选择一个 Algorithm (算法) 来应用于 training data set

algorithm 的选择取决于 模型试图解决的问题的 nature

3° Step 3: 训练 algorithm 来建立模型

是一个调整模型变量与系数以精准预测结果的过程.

无需 human intervention.

4° Step 4: 使用并提升 model

feed new data 来提升 effectiveness 与 accuracy

4. Supervised learning (监督学习)

Supervised machine learning 算法用 labeled data 作为 training data.

- 对任一 sample (x^i, y^i), $i=1, \cdots, N$, input data x^i 与其 output y^i 是已知的 (y^i 作为 label)
- 算法会比较自己预测的 outputs 与正确的 outputs 来计算 model accuracy 并优化模型系数 (找一个函数 h 使 y^i 接近 $h(x^i)$)

Supervised learning

Is There a Cat or a Dog?

cat

dog

Input/Feature: x ⟹ Output/Target: y

Goal: find the mapping from x to the correct y.

15

Supervised learning

Training data

y=1 (cat) y=0 (dog) y=1 (cat) …… y=0 (dog)

Learning algorithm (optimization involved)

Classifier $h: X \rightarrow \{0,1\}$

*Sometimes we also use $\{-1,1\}$ to denote the label

For example:

h  → 0

16

Supervised machine learning algorithms use **labeled data** as training data where the appropriate outputs to input data are known.

- For all samples, (x^i, y^i), i = 1, …N, you can observe both the input data x^i and the label y^i

Training data

y=1 (cat) y=0 (dog) y=1 (cat) …… y=0 (dog)

The algorithm compares its own predicted outputs with the correct outputs to calculate model accuracy and then optimizes model parameters to improve accuracy.

- Find a function h such that y^i is close to h(x^i)

§2 How to conduct supervised learning: K-Nearest neighbor classifier

1. Classification problem (分类问题)

Recap: Classification problems

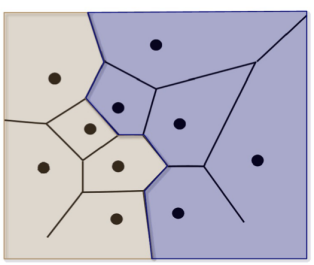
Samples:

- (x^i, y^i), i = 1, …N.
- y^i is the label which is discrete.

Target: predict the label y^{new} given the input data x^{new}

2. 方法一: K-Nearest neighbor classifier (K最邻近分类算法)

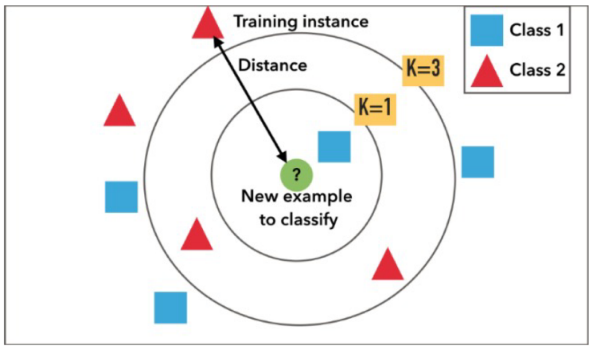
- 找K个最靠近 x 的 training points x_i.
- 若 K-nearest neighbours 中有较多的属于 classifier C, 则 label x 为 C.



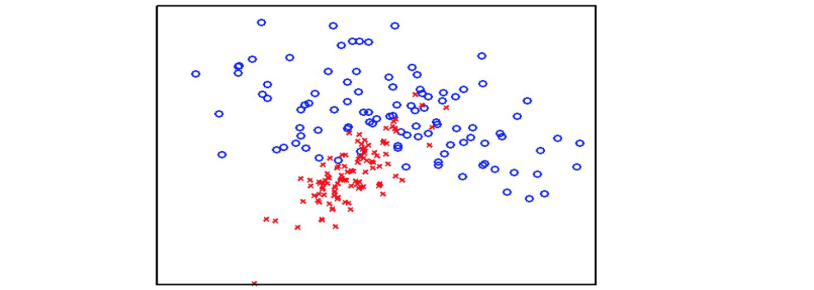
K=1

The KNN Algorithm

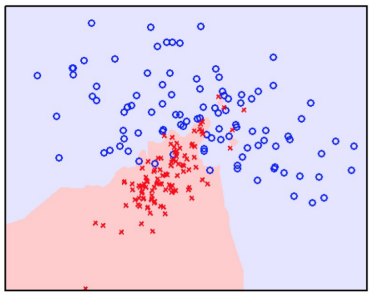
1. Load the data
2. Initialize K to your chosen number of neighbors
3. For each example in the data
   - Calculate the distance between the query example and the current example from the data.
   - Add the distance and the index of the example to an ordered collection
4. Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances
5. Pick the first K entries from the sorted collection
6. Get the labels of the selected K entries
7. Choose the label with the largest frequency



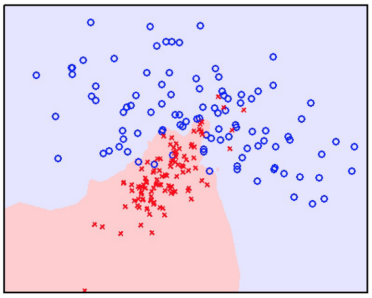
Example



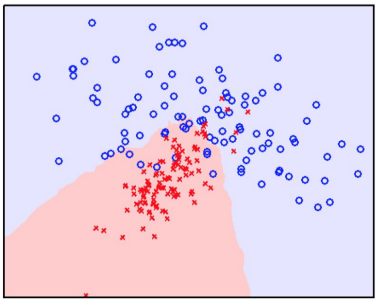
K = 1



K = 3



K = 5



K = 25

注：K的选取非常重要

K较小：noise对结果的影响较大

K较大：破坏KNN的基本原则：若K > 总data数，任一新input都有相同的label.

注：KNN的决策边界为非线性的

§3 How to conduct supervised learning: logistic regression model

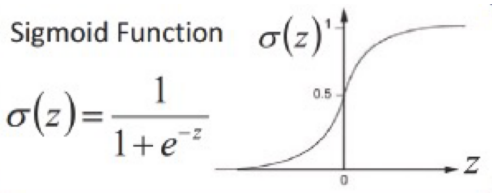
1. 方法二：logistic regression model（逻辑回归模型）

1° Model the (conditional) probability of the label for each feature

2° 用所有的 samples 来测算 conditional probability model.

2. Sigmoid function

$$\sigma(z) = \frac{1}{1+e^{-z}}$$



个人理解：

Sigmoid Function 是一个单调的函数，且是一个从 R 到 (0,1) 上的映射，对于一个 R 上的连续型变量 z，均可以找到一个对应的 $\sigma(z)$.

由于值域为 (0,1)，可以将 $\sigma(z)$ 视作一条件概率的值.

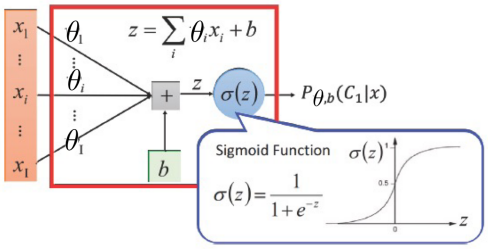
3. logistic regression model

逻辑回归处理的是分类问题！

- Simplest case (two classes): $y \in \{0,1\}$
- logistic regression model:

$$P(y=1|x,\theta,b) = \frac{1}{1+\exp(-(\theta^T x+b))}$$

$$P(y=0|x,\theta,b) = \frac{\exp(-(\theta^T x+b))}{1+\exp(-(\theta^T x+b))} = 1 - P(y=1|x,\theta,b)$$



对于一维的 x：

$$P(y=1|x,\theta,b) = \frac{1}{1+\exp(-(\theta x+b))}$$

对于多分类问题：$y \in \{0,1,\cdots\cdots,k\}$

$$\frac{P(y=i)}{P(y=0)} = e^{\theta_i^T x + b_i}$$

$$P(y=0|x,\theta,b) = \frac{\exp(-(\theta_i^T x+b_i))}{1+\sum_{i=1}^{k}\exp(-(\theta_i^T x+b_i))}$$

个人理解：

- 上式中的条件概率并非真实的概率，是用于分类的参考数值
- 以 classes = 2 为例，需要设定一个 threshold value（阈值）（通常为 1），
  - 若 $\frac{P(y=1)}{P(y=0)} = e^{\theta^T x+b} > 1$（此处默认阈值为 1），即 $P(y=1) > 0.5$，即 $\theta^T x+b > 0$ 时，x 对应的 y 取 1
  - 若 $\frac{P(y=1)}{P(y=0)} = e^{\theta^T x+b} < 1$（此处默认阈值为 1），即 $P(y=1) < 0.5$，即 $\theta^T x+b < 0$ 时，x 对应的 y 取 0
- 逻辑回归得出的决策边界为线性的，因此其适用于数据点可以用线性边界进行划分的情况。若要处理非线性分界的情况，可以将 $\theta^T x+b$ 换成非线性表达式。

4. MLE：找出合适的 θ 与 b

- 给出 labeled samples (x^i, y^i)，$i = 1, 2, \cdots\cdots, m$
- 找出 θ，使得 likelihood of the labels 最大

$$\max_\theta l(\theta, b) := \log \prod_{i=1}^m P(y^i | x^i, \theta, b) = \sum_{i=1}^m \log P(y^i | x^i, \theta, b)$$

- 通常，我们只需将 averaged likelihood 最大化

$$\max_\theta \frac{1}{m} l(\theta, b) := \frac{1}{m} \sum_{i=1}^m \log P(y^i | x^i, \theta, b)$$

1° Good news：$l(\theta, b)$ is concave in (θ, b)

- $$P(y=1 | x, \theta, b) = \frac{1}{1+\exp(-\theta x-b)}$$

$$P(y=0 | x, \theta, b) = \frac{\exp(-\theta x-b)}{1+\exp(-\theta x-b)}$$

由此可以得出：

$$\log P(y^i | x^i, \theta, b) = (y^i-1)(\theta x^i+b) - \log(1+\exp(-\theta x^i-b))$$

* 此式将 $y^i=1$ 与 $y^i=0$ 的情况均纳入其中

由此要证明 $l(\theta, b) := \sum_{i=1}^m \log P(y^i | x^i, \theta, b)$ concave，

仅需要证明 $-\log(1+\exp(-\theta x^i-b))$ concave，即 $\log(1+\exp(-\theta x^i-b))$ convex.

（若 $f_1, f_2, \cdots\cdots, f_n$ are convex functions，then $f_1+f_2+\cdots\cdots+f_n$ is also convex）

- 下面证明 $\log(1+\exp(-\theta x^i-b))$ is convex in (θ, b)
  - 引理：若 $f(\vec{x}+t\vec{a})$ 对任意 $\vec{x}$ 与 $\vec{a}$ 均 convex in t，则 f is convex in $\vec{x}$

证明：令 $g(t) = f(\vec{x}+t\vec{a})$，取 $\vec{x}_1, \vec{x}_2$，使 $\vec{x}_2 = \vec{x}_1 + t_1\vec{a}$

若 $f(\vec{x}+t\vec{a})$ is convex in t

则 $\lambda g(0) + (1-\lambda) g(t_1) \geq g((1-\lambda)t_1)$

即 $\lambda f(\vec{x}_1+0\vec{a}) + (1-\lambda) f(\vec{x}_1+t_1\vec{a}) \geq f(\vec{x}_1+(1-\lambda)t_1\vec{a})$

即 $\lambda f(\vec{x}_1) + (1-\lambda) f(\vec{x}_2) \geq f(\lambda\vec{x}_1+(1-\lambda)\vec{x}_2)$

$$= f(\lambda\vec{x}_1+(1-\lambda)(\vec{x}_1+t_1\vec{a}))$$

$$= f(\vec{x}_1+(1-\lambda)t_1\vec{a})$$

· 令 $(\theta, b) = (\theta_0, b_0) + t\vec{e}$，$(\theta_0, b_0)$ 为 any point，$\vec{e} = (\theta_1, b_1)$ 为任意单位向量

则 $h(t) := \log(1 + \exp(-\theta x - b))$

$= \log\{1 + \exp[(-\theta_0 x - b_0) + t(-\theta_1 x - b_1)]\}$

$= \log\{1 + \exp[C_1 + tC_2]\}$

则 $h''(t) = \frac{C_2^2}{(1+\exp[C_1+tC_2])^2} \cdot \exp[C_1 + tC_2] \geq 0$

2° Bad news: no closed form solution maximizing $l(\theta, b)$

$$\frac{1}{m} l(\theta, b) = \frac{1}{m} \sum_i \left((y^i - 1)(\theta x^i + b) - \log(1 + \exp(-\theta x^i - b))\right)$$

- Gradient (θ is one-dimension)

$$\frac{\partial l(\theta)}{\partial \theta} = \sum_i \left((y^i - 1)\, x^i + \frac{\exp(-\theta x^i - b) x^i}{1 + \exp(-\theta x^i - b)}\right)$$

$$\frac{\partial l(\theta)}{\partial b} = \sum_i \left((y^i - 1) + \frac{\exp(-\theta x^i - b)}{1 + \exp(-\theta x^i - b)}\right)$$

- Setting it to 0 does not lead to closed form solution

5. Numerical methods: 找出 (θ, b) 使 $l(\theta, b)$ 最大

1° Gradient descent (梯度下落法) 使 $f(x)$ 最小

· initial point 为 $x^{(0)}$

· 根据以下规则 update point:

$x^{(t+1)} = x^{(t)} - \alpha^{(t)} \cdot f'(x^{(t)})$

* $\alpha^{(t)}$ 为步长，$-f'(x^{(t)})$ 为负梯度，决定方向

· Stopping criteria:

$|x^{(t+1)} - x^{(t)}| \leq \varepsilon$

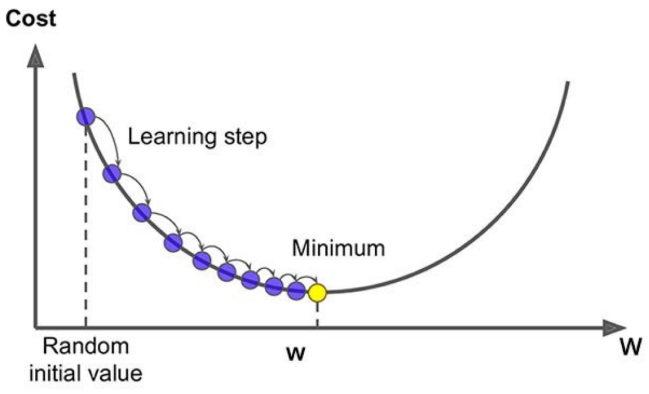
或 $|f'(x^{(t)})| \leq \varepsilon$

2° 步长 $\alpha^{(t)}$ 的选取

步长 $\alpha^{(t)}$ 的选取会影响找到极小值的速度，甚至是否能找到极小值.

· 前几步 $\alpha^{(t)}$ 可以大一些，靠近极小值时小一些

We may want $\alpha^{(t)}$ to be large for the first several steps but be small when approaching the local minimizer



- 若 $\alpha^{(t)}$ 为常数，可能会找不到极小值点

If $\alpha^{(t)}$ is a constant, we may not find a stable point.

f(x) = x^2

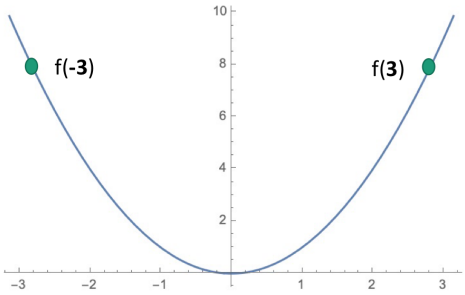
Suppose $\alpha^{(t)}$ = 1.

$$x^{(t+1)} = x^{(t)} - \alpha^{(t)} \nabla f(x^{(t)})$$

$$x^{(t+1)} = x^{(t)} - f'(x^{(t)})$$

If $x^{(t)}$ = -3
- f'(-3) = -6
- $x^{(t+1)}$ = 3

If $x^{(t)}$ = 3
- f'(3) = 6
- $x^{(t+1)}$ = -3



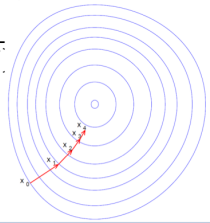
The updated solution will always either be 3 or -3

3° 用于逻辑回归的梯度下落算法

- Initialize parameter (θ^0,b^0)
- Do

$$\theta^{t+1} \leftarrow \theta^t + \alpha^{(t)} \frac{1}{m} \sum_i (y^i - 1)\, x^i + \frac{\exp(-\theta^t x^i - b^t) x^i}{1 + \exp(-\theta^t x^i - b^t)}$$

$$b^{t+1} \leftarrow b^t + \alpha^{(t)} \frac{1}{m} \sum_i (y^i - 1) + \frac{\exp(-\theta^t x^i - b^t)}{1 + \exp(-\theta^t x^i - b^t)}$$



- While $|\theta^{t+1} - \theta^t| > \epsilon$ or $|b^{t+1} - b^t| > \epsilon$

$\alpha^{(t)}$: the step size or learning rate

4° A variant: Stochastic gradient descent (随机梯度下落)

- At each iteration, we randomly choose a small batch of data points, and update using the stochastic gradient

$$\theta^{t+1} \leftarrow \theta^t + \alpha^{(t)} \frac{1}{|B|} \sum_{i \in B} (y^i - 1)\, x^i + \frac{\exp(-\theta^t x^i - b^t) x^i}{1 + \exp(-\theta^t x^i - b^t)}$$

$$b^{t+1} \leftarrow b^t + \alpha^{(t)} \frac{1}{|B|} \sum_{i \in B} (y^i - 1) + \frac{\exp(-\theta^t x^i - b^t)}{1 + \exp(-\theta^t x^i - b^t)}$$

- B: the batch we use in each iteration