Gradient Descent

Brief review of most basic optimisation approach in ML

Optimisation formulations in ML

- Training = Fitting = Parameter estimation
- Typical formulation → objective function w r t data and Θ

$$\hat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}}\, L(data, \boldsymbol{\theta})$$

constrain if no constrain Θ is a vector ~~with~~ of dimension p

  - argmin because we want a minimiser not the minimum
    - Note: argmin can return a set (minimiser not always unique!)
  - Θ denotes a model family (including constraints)
  - L denotes some objective function to be optimised
    - E.g. MLE: (conditional) likelihood log-likelihood for linear
    - E.g. Decision theory: (regularised) empirical risk

One we've seen: Log trick

optimize $L(\theta)$ = optimize $\log L(\theta)$

- Instead of optimising $L(\theta)$, try convenient $\log L(\theta)$

quizzes.

- Why are we allowed to do this?

ans:

- Strictly monotonic function: a $> b \Longrightarrow f(a) > f(b)$
  - Example: log function!

Two solution approaches

- Analytic (*aka* closed form) solution
  - Known only in limited number of cases
  - Use 1st-order necessary condition for optimality*: gradient
    $$\frac{\partial L}{\partial \theta_1} = \cdots = \frac{\partial L}{\partial \theta_p} = 0$$

Assuming unconstrained, differentiable *L*

LR都可

- Approximate iterative solution Logistic regression only.
  1. Initialisation: choose starting guess $\boldsymbol{\theta}^{(1)}$, set $i = 1$
  2. Update: $\boldsymbol{\theta}^{(i+1)} \leftarrow SomeRule\left[\boldsymbol{\theta}^{(i)}\right]$, set $i \leftarrow i + 1$
  3. Termination: decide whether to Stop
  4. Go to Step 2
  5. Stop: return $\hat{\boldsymbol{\theta}} \approx \boldsymbol{\theta}^{(i)}$

depend on previous point

e.g., $\theta \to 2\theta \to 3\theta$
选择

* Note: to check for local minimum, need positive 2nd derivative (or Hessian positive definite); this assumes unconstrained – in general need to also check boundaries. See also Lagrangian techniques later in subject.

Reminder: The gradient

- Gradient at $\boldsymbol{\theta}$ defined as $\left[\frac{\partial L}{\partial \theta_1}, \dots, \frac{\partial L}{\partial \theta_p}\right]'$ evaluated at $\boldsymbol{\theta}$
- The gradient points to the direction of maximal change of $L(\boldsymbol{\theta})$ when departing from point $\boldsymbol{\theta}$
- Shorthand notation
  * $\boldsymbol{\nabla} L \stackrel{\text{def}}{=} \left[\frac{\partial L}{\partial \theta_1}, \dots, \frac{\partial L}{\partial \theta_p}\right]'$ computed at point $\boldsymbol{\theta}$
  * Here $\boldsymbol{\nabla}$ is the "nabla" symbol

Harps, p. 984.

- Hessian matrix $\boldsymbol{\nabla}^2 L$ at $\boldsymbol{\theta}$: $(\boldsymbol{\nabla}^2 L)_{ij} = \frac{\partial^2 L}{\partial \theta_i \partial \theta_j}$

二阶导矩阵

Gradient descent and SGD

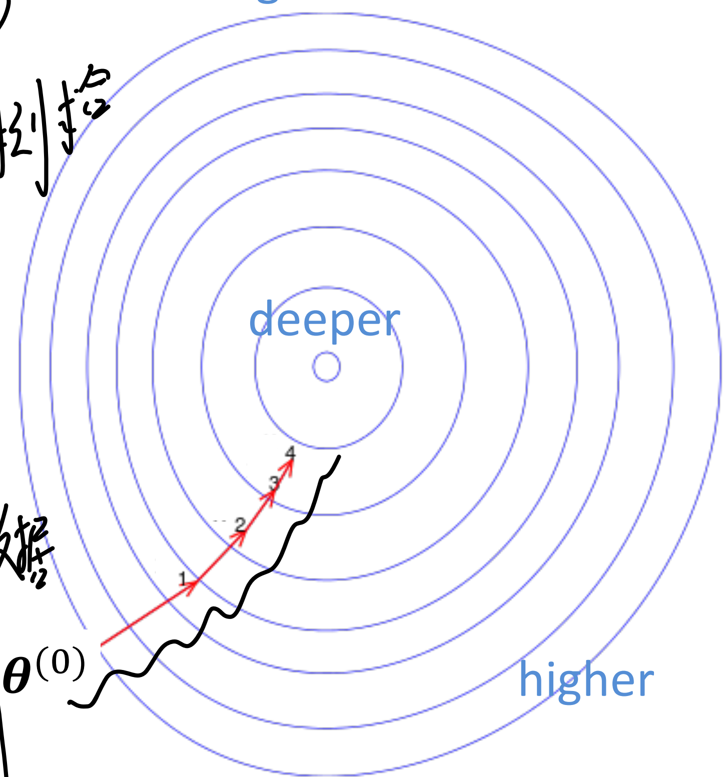
Assuming L is differentiable

1. Choose $\boldsymbol{\theta}^{(1)}$ and some T
2. For i from 1 to T*
   1. $\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \eta \nabla L(\boldsymbol{\theta}^{(i)})$ → 2阶导数要大于0

      每个数据点(这一步)的梯度
3. Return $\widehat{\boldsymbol{\theta}} \approx \boldsymbol{\theta}^{(i)}$

- Note: η dynamically updated per step
- Variants: Momentum, AdaGrad, …
- Stochastic gradient descent: two loops 选取部分数据
  - Outer for loop: each loop (called epoch) sweeps through all training data
  - Within each epoch, randomly shuffle training data; then for loop: do gradient steps only on batches of data. Batch size might be 1 or few

Viewing L from above:



Wikimedia Commons. Authors: Olegalexandrov, Zerodamage

*Other stopping criteria can be used

每个batch下降 $\theta^1 \to \theta^2$ 用batch1; $\theta^2 \to \theta^3$ 用batch2

耗时少，计算量小

SGD ⇒ batch里用一个点

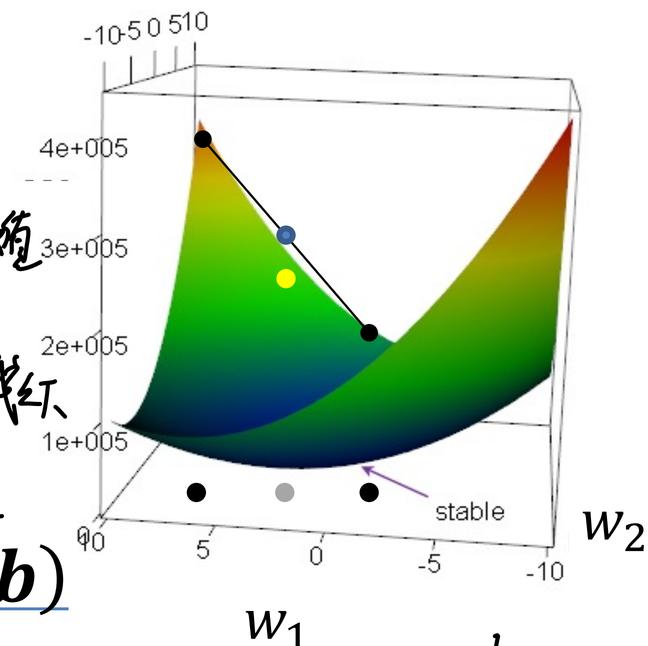
Convex objective functions

- '**Bowl shaped**' functions
- Informally: if line segment between any two points on graph of function lies above or on graph

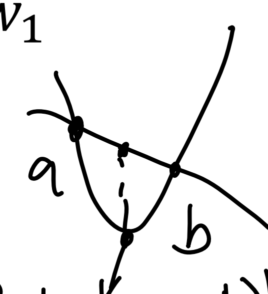
函数值永远在他们线下

- Formally* $f: D \to \mathbf{R}$ is convex if $\forall \boldsymbol{a}, \boldsymbol{b} \in D, t \in [0,1]$:
  $f(t\boldsymbol{a} + (1-t)\boldsymbol{b}) \leq tf(\boldsymbol{a}) + (1-t)f(\boldsymbol{b})$
  Strictly convex if inequality is strict (<)

凸函数



- **Gradient descent** on (strictly) convex function guaranteed to find a (unique) global minimum!



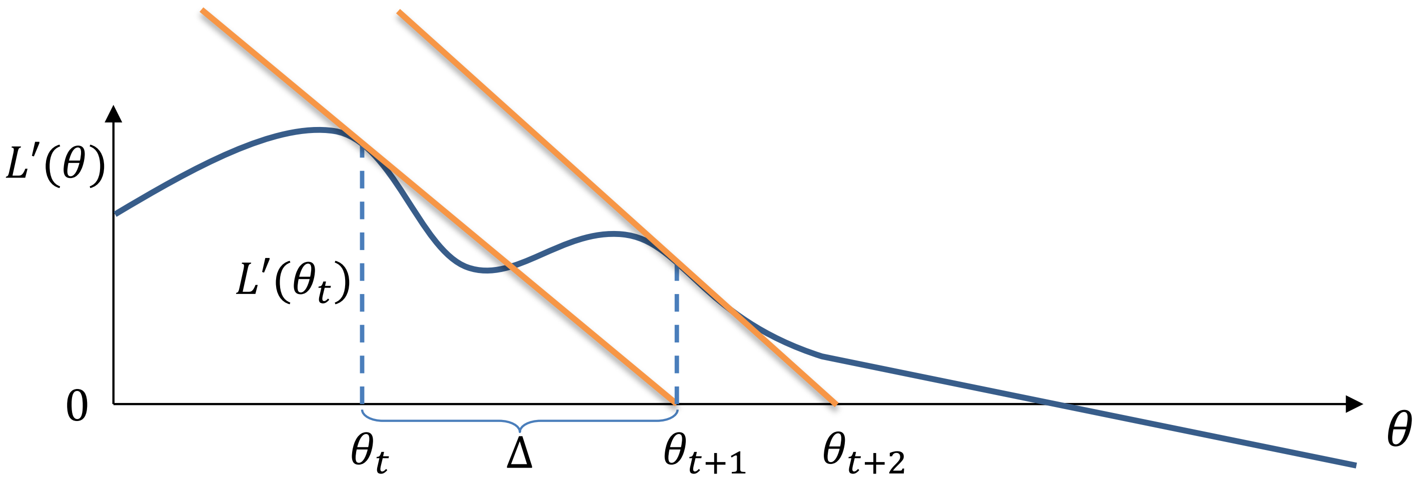
凸函数能找到 global minimum

* Aside: Equivalently we can look to the second derivative. For f defined on scalars, it should be non-negative; for multivariate f, the Hessian matrix should be positive semi-definite (see linear algebra supplemental deck).

Newton-Raphson

A second-order method;
Successive root finding in the objective's derivative.

Newton-Raphson: Derivation (1D)

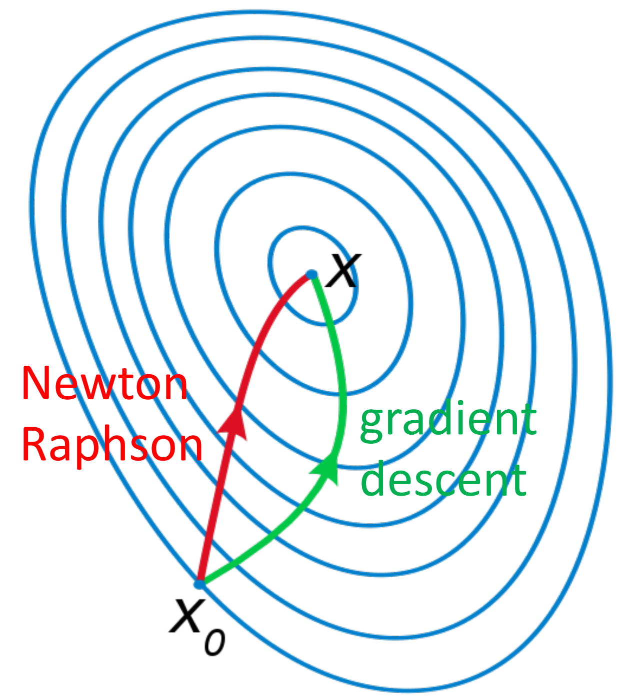


- Critical points of $L(\theta)$ = Zero-crossings of $L'(\theta)$
- Consider scalar θ. Starting at given/arbitrary θ_0, iteratively:
  1. Fit tangent line to $L'(\theta)$ at θ_t
  2. Need to find $\theta_{t+1} = \theta_t + \Delta$ using linear approximation's zero crossing
  3. Tangent line given by derivative: rise/run = $-L''(\theta_t) = L'(\theta_t)/\Delta$
  4. Therefore iterate is $\theta_{t+1} = \theta_t - L'(\theta_t)/L''(\theta_t)$ 二阶导 SGD的 前面是(学习率×一阶)

https://math.stackexchange.com/questions/4339706/why-does-the-sign-in-newtons-method-matter

Newton-Raphson: General case

- Newton-Raphson summary
  * Finds $L'(\theta)$ zero-crossings
  * By successive linear approximations to $L'(\theta)$
  * Linear approximations involve derivative of $L'(\theta)$, ie. $L''(\theta)$



public domain wikipedia

- Vector-valued θ:
  How to fix scalar $\theta_{t+1} = \theta_t - L'(\theta_t)/L''(\theta_t)$???
  * $L'(\theta)$ is $\nabla L(\theta)$
  * $L''(\theta)$ is $\nabla^2 L(\theta)$
  * Matrix division is matrix inversion

逆矩阵

二阶导计算消耗大

- General case: $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \left(\nabla^2 L(\boldsymbol{\theta}_t)\right)^{-1} \nabla L(\boldsymbol{\theta}_t)$
  * Pro: May converge faster; fitting a quadratic with curvature information
  * Con: Sometimes computationally expensive, unless approximating Hessian

Conclusion:

梯度下降方法: { 一阶 GD, 二阶 Newton }

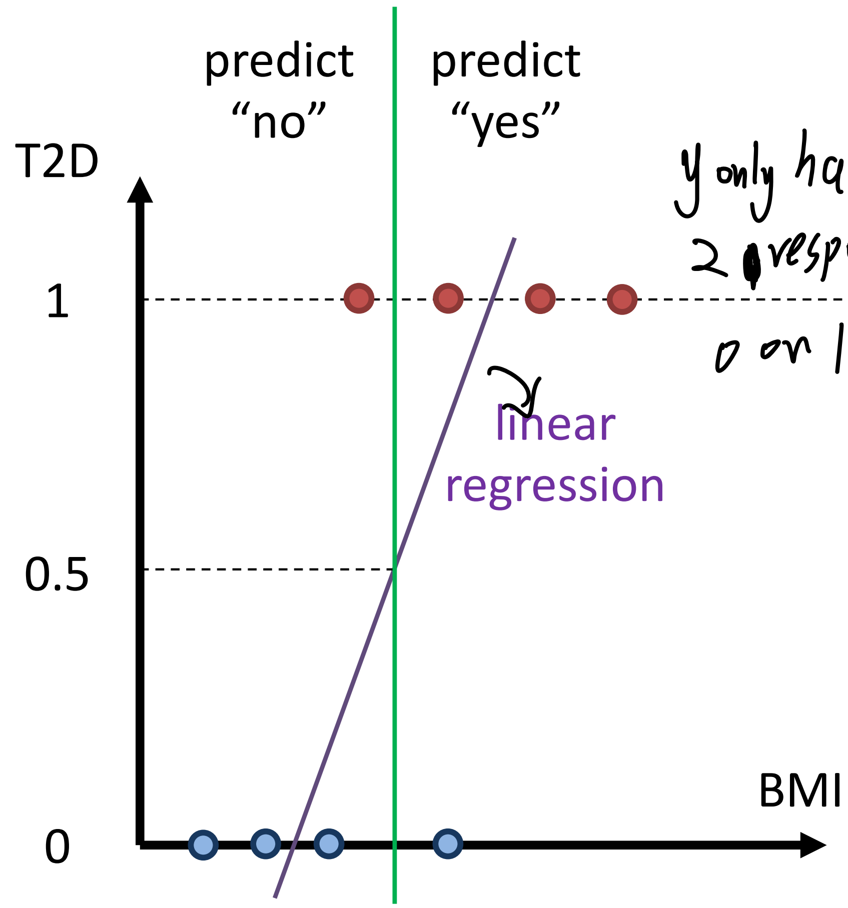
{ Batch GD, SGD }

Logistic Regression Model

A workhorse linear, binary classifier;
(A review for some of you; new to some.)

Binary classification: Example

- Example: given body mass index (BMI) does a patient have type 2 diabetes (T2D)?
- This is (supervised) binary classification
- One *could* use linear regression
  - Fit a line/hyperplane to data (find weights $\boldsymbol{w}$)
  - Denote $s \equiv \boldsymbol{x}'\boldsymbol{w}$
  - Predict "Yes" if $s \geq 0.5$
  - Predict "No" if $s < 0.5$



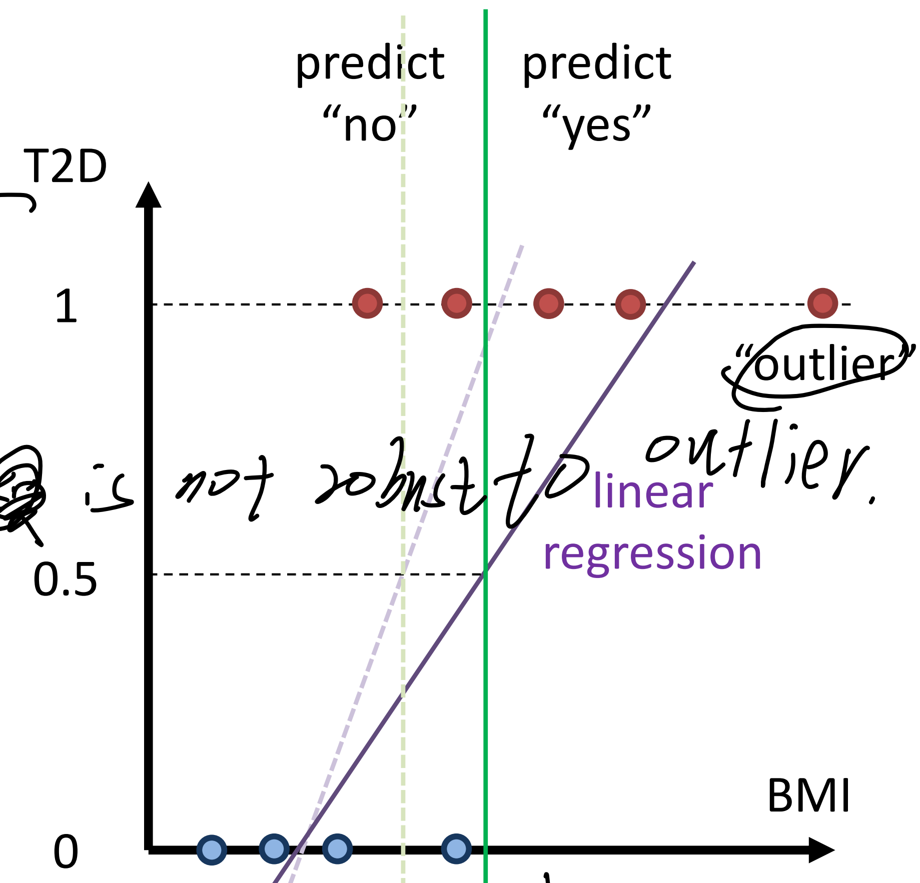
y only has 2 responses
0 or 1

Why not linear regression

- Due to the square loss, points far from boundary have loss squared – even if they're confidently correct!

least square is not robust to outlier.

- Such "outliers" will "pull at" the linear regression
- Overall, the least-squares criterion looks unnatural in this setting



★LR:适合连续值的预测.

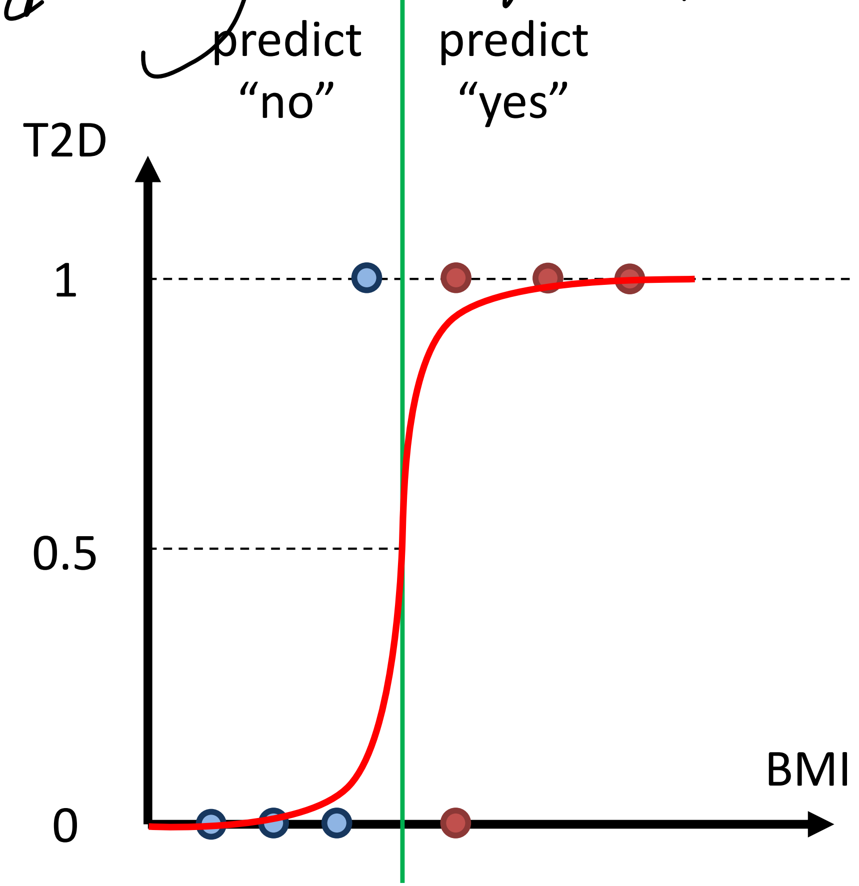
★outlier的 squared error非常大导致决策边界的偏移(扭曲)，会导致其他点无法被正确分类

Logistic regression model

less affected by outlier.

- Probabilistic approach to classification
  - $P(Y=1|\boldsymbol{x}) = f(\boldsymbol{x}) = ?$
  - Use a linear function? E.g., $s(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{w}$
- Problem: the probability needs to be between 0 and 1.
- Logistic function $f(s) = \frac{1}{1+\exp(-s)}$
- Logistic regression model

$$P(Y=1|\boldsymbol{x}) = \frac{1}{1+\exp(-\boldsymbol{x}'\boldsymbol{w})}$$



How is logistic regression *linear*?

- Logistic regression model:

$$P(Y=1|\boldsymbol{x}) = \frac{1}{1+\exp(-\boldsymbol{x}'\boldsymbol{w})}$$

- Classification rule:

$\le \frac{1}{2}$ 就是0

if $\left(P(Y=1|\boldsymbol{x}) > \frac{1}{2}\right)$ then class "1", else class "0"

- Decision boundary is the set of **x**'s such that:

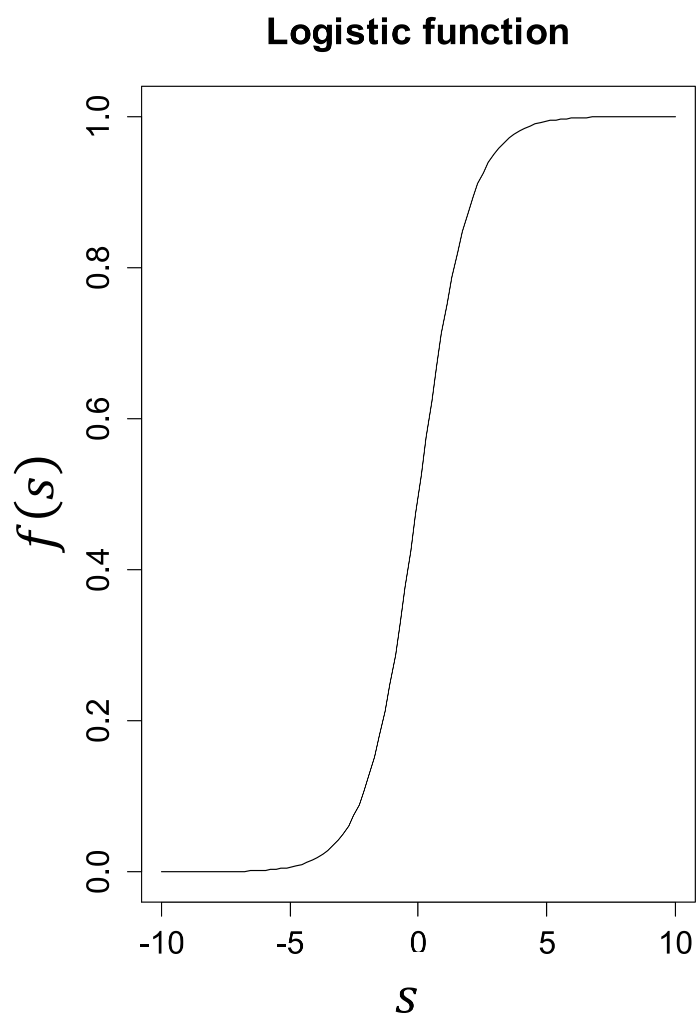
Decision boundary:

$$\frac{1}{1+\exp(-\boldsymbol{x}'\boldsymbol{w})} = \frac{1}{2}$$

Logistic Regression

$$\exp(-\boldsymbol{x}'\boldsymbol{w}) = 1$$

$$\boldsymbol{x}'\boldsymbol{w} = 0$$



Linear vs. logistic probabilistic models

- Linear regression assumes a Normal distribution with a fixed variance and mean given by linear model

$$p(y|\boldsymbol{x}) = Normal(\boldsymbol{x}'\boldsymbol{w}, \sigma^2)$$

线性回归是 正态

- Logistic regression assumes a Bernoulli distribution with parameter given by logistic transform of linear model

$$p(y|\boldsymbol{x}) = Bernoulli(\text{logistic}(\boldsymbol{x}'\boldsymbol{w}))$$

Logistic是Bernoulli

- Recall that Bernoulli distribution is defined as

$$p(1) = \theta \text{ and } p(0) = 1 - \theta \text{ for } \theta \in [0,1]$$

- Equivalently $p(y) = \theta^y(1-\theta)^{(1-y)}$ for $y \in \{0,1\}$

Training as Max-Likelihood Estimation

- Assuming independence, probability of data

$$p(y_1, \dots, y_n | \boldsymbol{x}_1, \dots, \boldsymbol{x}_n) = \prod_{i=1}^{n} p(y_i | \boldsymbol{x}_i)$$

- Assuming Bernoulli distribution we have

$$p(y_i | \boldsymbol{x}_i) = \left(\theta(\boldsymbol{x}_i)\right)^{y_i} \left(1 - \theta(\boldsymbol{x}_i)\right)^{1-y_i}$$

where $\theta(\boldsymbol{x}_i) = \frac{1}{1+\exp(-\boldsymbol{x}_i{}'\boldsymbol{w})}$

w是一样的

- Training: maximise this expression wrt weights $\boldsymbol{w}$

Apply log trick, simplify

- Instead of maximising likelihood, maximise its logarithm

$$\log\left(\prod_{i=1}^{n} p(y_i|\boldsymbol{x_i})\right) = \sum_{i=1}^{n} \log p(y_i|\boldsymbol{x_i})$$

$$= \sum_{i=1}^{n} \log\left(\left(\theta(\boldsymbol{x}_i)\right)^{y_i}\left(1-\theta(\boldsymbol{x}_i)\right)^{1-y_i}\right)$$

$$= \sum_{i=1}^{n}\left(y_i \log\left(\theta(\boldsymbol{x}_i)\right) + (1-y_i)\log\left(1-\theta(\boldsymbol{x}_i)\right)\right)$$

$$\theta(x_i) = \frac{1}{1+\exp(-x_i' x)}$$

$$= \sum_{i=1}^{n}\left((y_i - 1)\boldsymbol{x_i'w} - \log(1+\exp(-\boldsymbol{x_i'w}))\right)$$

凸函数解不了,但可用梯度下降搞

Can't do this analytically

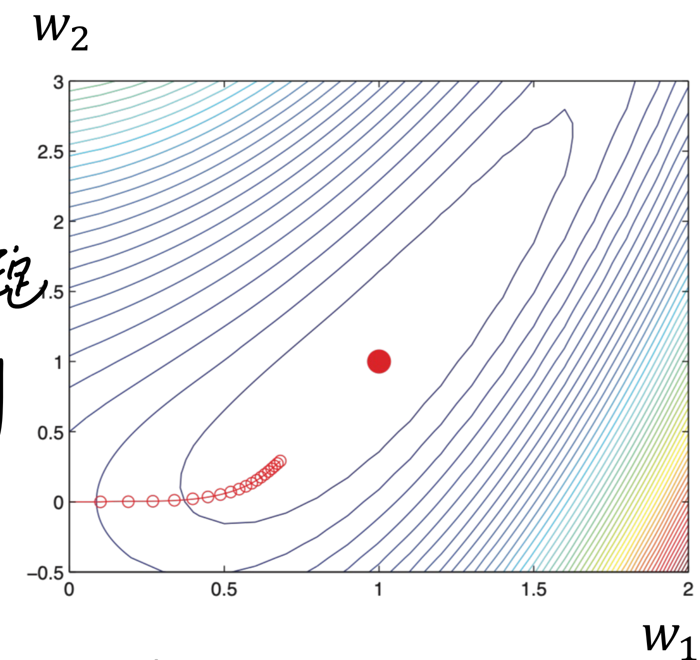
Training as Iterative Optimisation

- Training logistic regression: $\mathbf{w}$ maximising log-likelihood $L(\mathbf{w})$ or cross-entropy loss
- **Bad news**: No closed form solution
- **Good news**: Problem is strictly convex, if no irrelevant features → convergence!

→有逆矩阵

x_1 x_2 x_3 特征 不能是 linear combination $(X^T X)^{-1} X^T y$

Look ahead: regularisation for irrelevant features



Murphy, Fig 8.3, p247

How does gradient descent work?

- simply take gradient of log-likelihood, i.e.,
  $\nabla L(\mathbf{w}) = \sum_{i=1}^{n} \left(y_i - \theta(\mathbf{x}_i)\right)\mathbf{x}_i$
- plug into favourite iterative optimiser GD/SGD/Adagrad/Adam/BFGS/…

Background: Cross entropy

- Cross entropy is an information-theoretic method for comparing two distributions (reference dist VS estimated dist)
- Cross entropy is a measure of a divergence between reference distribution $g_{ref}(a)$ and estimated distribution $g_{est}(a)$. For discrete distributions:

$$H(g_{ref}, g_{est}) = -\sum_{a \in A} g_{ref}(a) \log g_{est}(a)$$

后面是est

A is support of the distributions, e.g., $A = \{0,1\}$

Training as cross-entropy minimisation

- Consider log-likelihood for a single data point
  $$\log p(y_i|\boldsymbol{x}_i) = y_i \log\big(\theta(\boldsymbol{x}_i)\big) + (1 - y_i) \log\big(1 - \theta(\boldsymbol{x}_i)\big)$$
- Cross entropy $H\left(g_{ref}, g_{est}\right) = -\sum_a g_{ref}(a) \log g_{est}(a)$
  * If reference (true) distribution is
    $$g_{ref}(1) = y_i \text{ and } g_{ref}(0) = 1 - y_i$$
  * With logistic regression estimating this distribution as
    $$g_{est}(1) = \theta(\boldsymbol{x}_i) \text{ and } g_{est}(0) = 1 - \theta(\boldsymbol{x}_i)$$

It finds $\boldsymbol{w}$ that minimises sum of cross entropies per training point