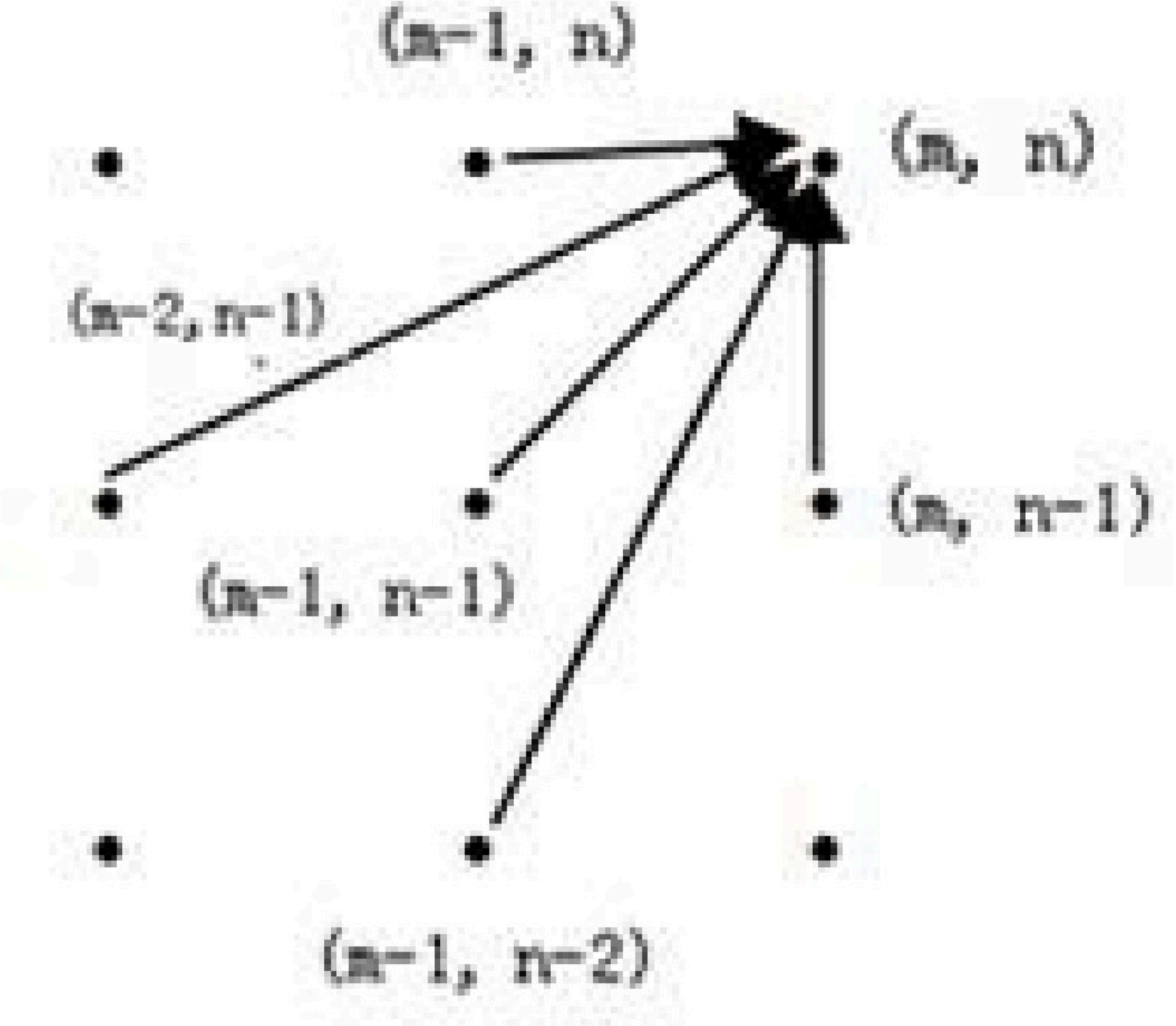
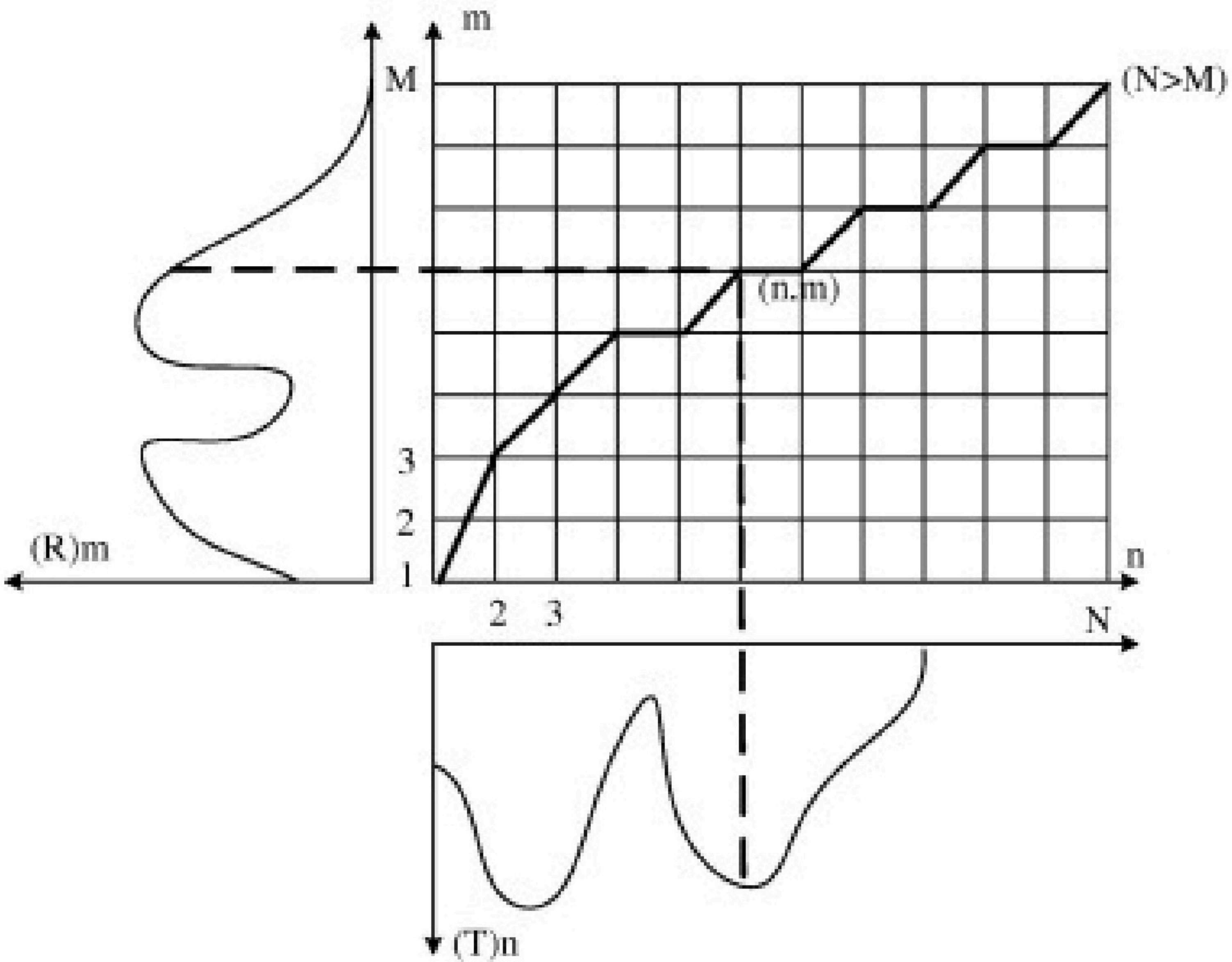
DTW
(Dynamic Time Warping)

计算两个序列之间的距离，比如可以用来做语音识别、股票相似K线的计算，任意不等长序列之间的距离计算。这个问题也叫做sequential alignment.

Applications


(m-1, n)
(m, n)
(m-2, n-1)
(m-1, n-1)
(m, n-1)
(m-1, n-2)


m
M
(N>M)
(n,m)
3
2
1
(R)m
n
2 3
N
(T)n

逻辑回归

Classification Problem

年龄	工资	学历	逾期
20	4000	本科	YES
25	5000	专科	NO
21	6000	本科	NO
25	5000	专科	YES
28	8000	本科	NO
27	7000	本科	$\boxed{?}$

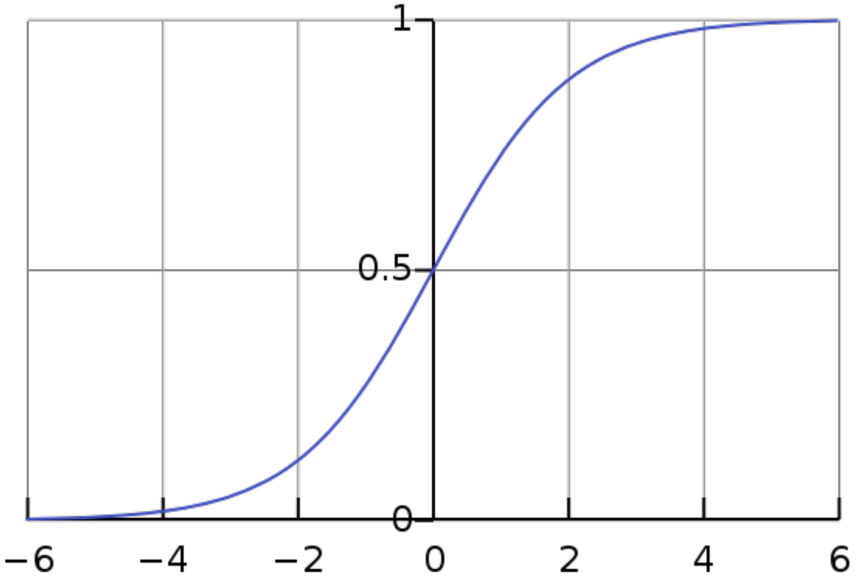
- **学习输入到输出的映射 $f: X \rightarrow Y$**
- X：输入
- Y：输出

Classification Problem

- 怎么去表示 $P(Y|X)$? 这相当于用模型来捕获输入X和输出Y之间的关系

- **可不可以用线性回归来表示$P(Y|X) = w^T x + b$? 为什么?**

Logistic Function



$$y = \frac{1}{1+e^{-x}}$$

$x: (-\infty, +\infty)$

$y: (0,1)$

可不可以把线性回归 $w^T x + b$ 改进一下使得值域映射到 (0，1)区间?

Logistic Function

逻辑函数 $y=\frac{1}{1+e^{-x}}$

原始条件概率：$P(Y|X)=w^{\mathrm{T}}x+b$

新条件概率 $P(Y|X)=\frac{1}{1+e^{-w^{\mathrm{T}}x+b}}$

年龄	工资	学历	逾期
20	4000	本科	YES
25	5000	专科	NO
21	6000	本科	NO
25	5000	专科	YES
28	8000	本科	NO
27	7000	本科	$\boxed{?}$

Logistic Function

对于二分类问题：

$$p(y=1|x,w)=\frac{1}{1+e^{-w^Tx+b}}$$

$$p(y=0|x,w)=\frac{e^{-w^Tx+b}}{1+e^{-w^Tx+b}}$$

两个式子可以合并成：

$$p(y|x,w)=p(y=1|x,w)^y[1-p(y=1|x,w)]^{1-y}$$

Logistic Regression is Linear Classifier

$$p(y = 1|x, w) = \frac{1}{1 + e^{-w^T x + b}}$$

$$p(y = 0|x, w) = \frac{e^{-w^T x + b}}{1 + e^{-w^T x + b}}$$

Objective Function

假设我们拥有数据集 $D = \{(x_i, y_i)\}_{i=1}^{n} \quad x_i \in R^d, \quad y_i \in \{0，\ 1\}$

而且我们已经定义了：

$$\boxed{p(y|x,w) = p(y=1|x,w)^y[1-p(y=1|x,w)]^{1-y}}$$

我们需要最大化目标函数：

$$\boxed{\widehat{w}_{MLE},\ \ \hat{b}_{MLE} = argmax_w \prod_{i=1}^{n} p(y_i|x_i, w, b)}$$

Objective Function

我们需要最大化目标函数：

$$\widehat{w}_{MLE},\quad \hat{b}_{MLE} = argmax_{w,b} \prod_{i=1}^{n} p(y_i|x_i, w, b)$$

Objective Function

$$argmin_{w,b} - \sum_{i=1}^{n} \log p(y_i | x_i, w, b)$$

$$\boxed{p(y|x,w) = p(y=1|x,w)^y [1 - p(y=1|x,w)]^{1-y}}$$

Minimizing the Function

求使得$f(w)$值最小的参数w

- **是否凸函数**
- **最优化算法**

Gradient Descent

求使得$f(w)$值最小的参数w

$$\boxed{\begin{array}{l} \text{初始化} w^1 \\ for\ t = 1,2,\ldots: \\ \quad w^{t+1} = w^t - \eta \nabla f(w^t) \end{array}}$$

例子： 求解函数 $f(w) = 4w^2 + 5w + 1$的最优解

Gradient Descent for Logistic Regression

$$p(y = 1|x, w) = \frac{1}{1 + e^{-w^{\mathrm{T}}x + b}}$$

$$argmin_{w,b} - \sum_{i=1}^{n} y \log p(y = 1|x, w) + (1 - y) \log(1 - p(y = 1|x, w)]$$

Gradient Descent for Logistic Regression

Stochastic Gradient Descent for Logistic Regression

Stochastic Gradient Descent for Logistic Regression

加入正则– L2 Norm

$$\hat{w}_{MLE}, \quad \hat{b}_{MLE} = argmin_{w,b} - \sum_{i=1}^{n} \log p(y_i|x_i, w, b) + \lambda||\mathrm{w}||_2^2$$

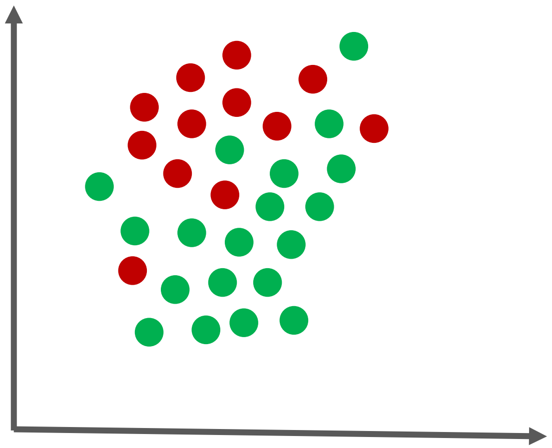
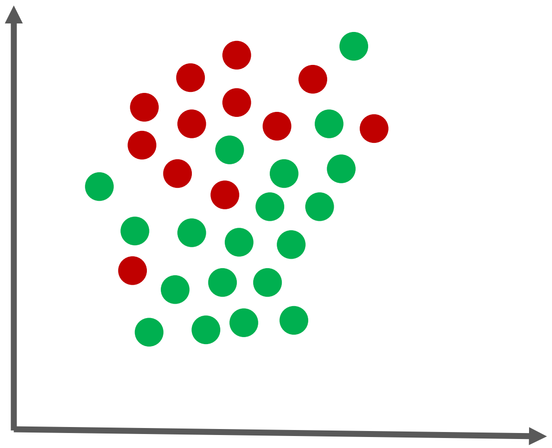
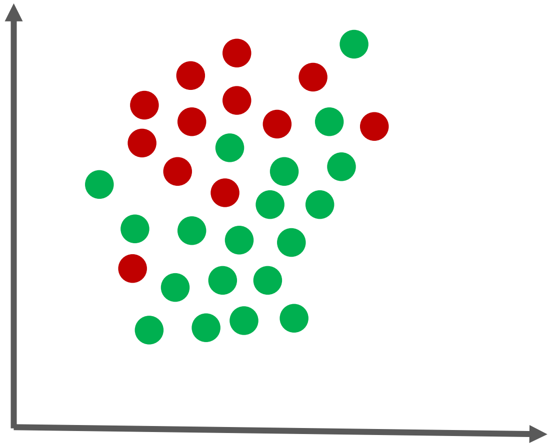
Gradient Descent

$$\widehat{w}_{MLE}, \quad \hat{b}_{MLE} = argmin_{w,b} - \sum_{i=1}^{n} \log p(y_i | x_i, w, b) + \lambda ||\mathrm{w}||_2^2$$

Stochastic Gradient Descent

$$\widehat{w}_{MLE}, \quad \widehat{b}_{MLE} = argmin_{w,b} - \sum_{i=1}^{n} \log p(y_i | x_i, w, b) + \lambda ||\mathrm{w}||_2^2$$

Model Complexity

Generalization Capability

Regularized Objective Function

Commonly Used Regularization Terms

L1

L2

Geometric Interpretation of L1 vs L2

Some Applications of L1 Regularization

How to Select λ?

$$\hat{w}_{MLE}, \quad \hat{b}_{MLE} = argmax_{w,b} \prod_{i=1}^{n} p(y_i | x_i, , w) + \lambda ||\mathrm{w}||_2^2$$

Model Parameter vs Hyperparameter

$$\hat{w}_{MLE}, \quad \hat{b}_{MLE} = argmax_{w,b} \prod_{i=1}^{n} p(y_i | x_i, w, b) + \lambda||\mathrm{w}||_2^2$$

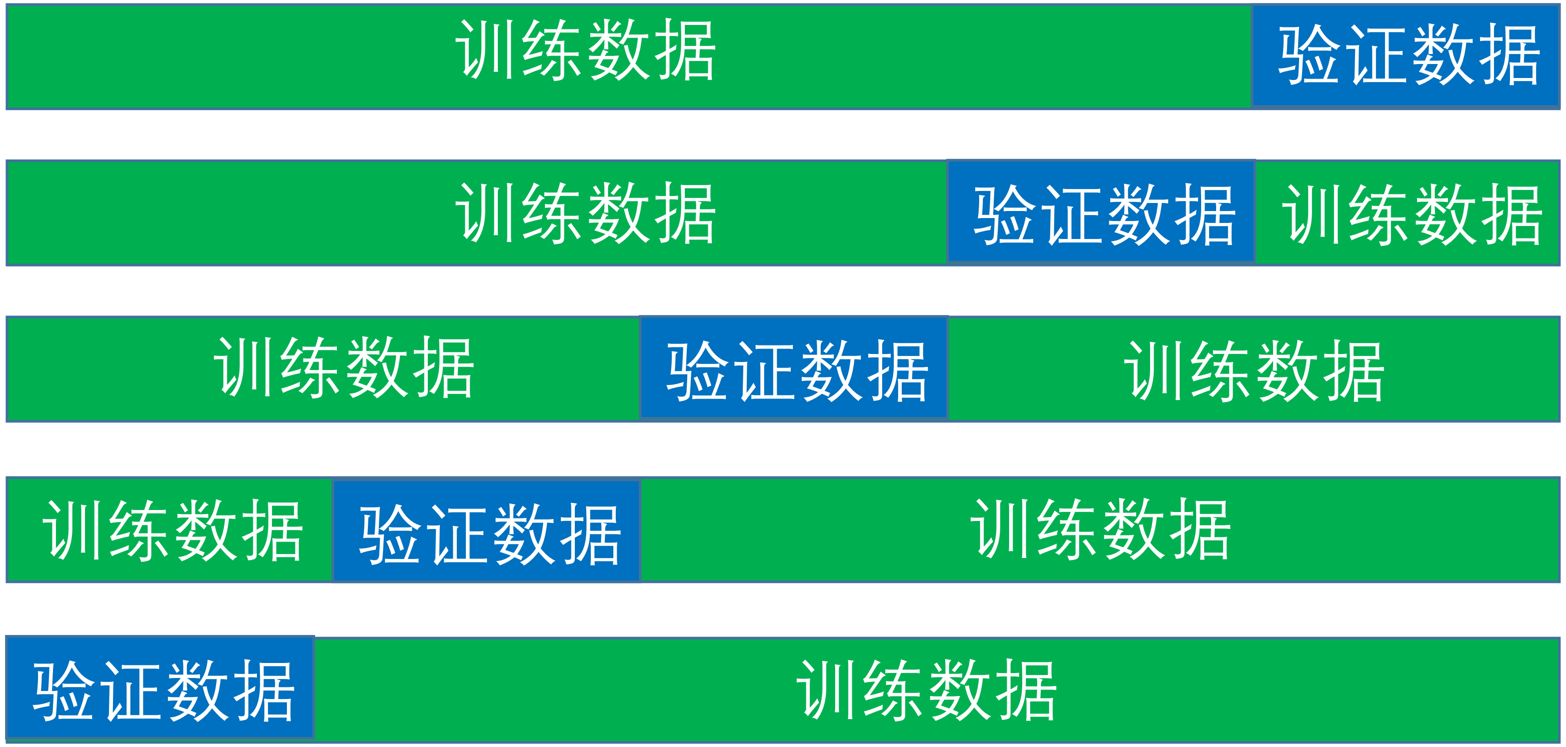
Cross Validation - Selecting Hyperparameter

Intuition

把训练数据进一步分成训练数据（Training Data）和验证集(Validation Data)。 选择在验证数据里最好的超参数组合。

5-Fold Cross Validation

测试数据



Very Important Note!

绝对不能用测试数据来引导（guide）模型的训练！

绝对不能用测试数据来引导（guide）模型的训练！

绝对不能用测试数据来引导（guide）模型的训练！

L1+L2 Regularized Logistic Regression

$$L = \prod_{i=1}^{n} p(y_i|x_i, w, b) + \lambda_1 ||\mathrm{w}||_1 \ + \lambda_2 ||\mathrm{w}||_2^2$$

How to search for λ_1 and λ_2 ?

Grid Search

$$L = \prod_{i=1}^{n} p(y_i|x_i, w, b) + \lambda_1 ||\mathrm{w}||_1 \ + \lambda_2 ||\mathrm{w}||_2^2$$

Heuristic Search

- 随机搜索（Random Search）
- 遗传算法（Genetic/Evolutionary Algorithm）
- 贝叶斯优化（Bayesian Optimization）

Neuro-Science

Time-Aware Recommendation

Summary

- 好的模型拥有高的泛化能力
- 越复杂的模型越容易过拟合
- 添加正则项是防止过拟合的一种手段
- L1正则会带来系数特性
- 选择超参数时使用交叉验证
- 参数搜索过程最耗费资源