人工智能

——机器学习基础II：支撑向量机

COMP130207.01

部分Slides来源于斯坦福课程CS221

线性模型：支撑向量机

背景知识

Norm范数的定义

- Assigns a positive number to each non-zero vector
- Is only zero if the vector is an all-zero vector
- Key aspect in proving uniqueness results

Norm properties

- Homogeneity: $\|\alpha \mathbf{x}\| = |\alpha| \|\mathbf{x}\|$, for $\mathbf{x} \in \mathbb{R}^N$ and $\alpha \in \mathbb{R}$
- Subadditivity: $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$, for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$
- Separability: If and only if $\|\mathbf{x}\| = 0$, then $\mathbf{x} = 0$

背景知识

Norm范数的定义

ℓ_p norms

Definition (ℓ_p norms)

$$p \geq 1, \mathbf{a} \in \mathbb{R}^N, \|\mathbf{a}\|_p = \left(\sum_{i=1}^{N} |a_i|^p \right)^{1/p}$$

- ℓ_2 norm: $p = 2, \|\mathbf{a}\|_2 = \sqrt{\sum_i |a_i|^2}$
- ℓ_1 norm: $p = 1, \|\mathbf{a}\|_1 = \sum_i |a_i|$
- ℓ_∞ norm: $p = \infty, \|\mathbf{a}\|_\infty = \max_i |a_i|$

Lemma (Minkowski's inequality)

$$1 \leq p \leq \infty, \quad \|\mathbf{a} + \mathbf{b}\|_p \leq \|\mathbf{a}\|_p + \|\mathbf{b}\|_p$$

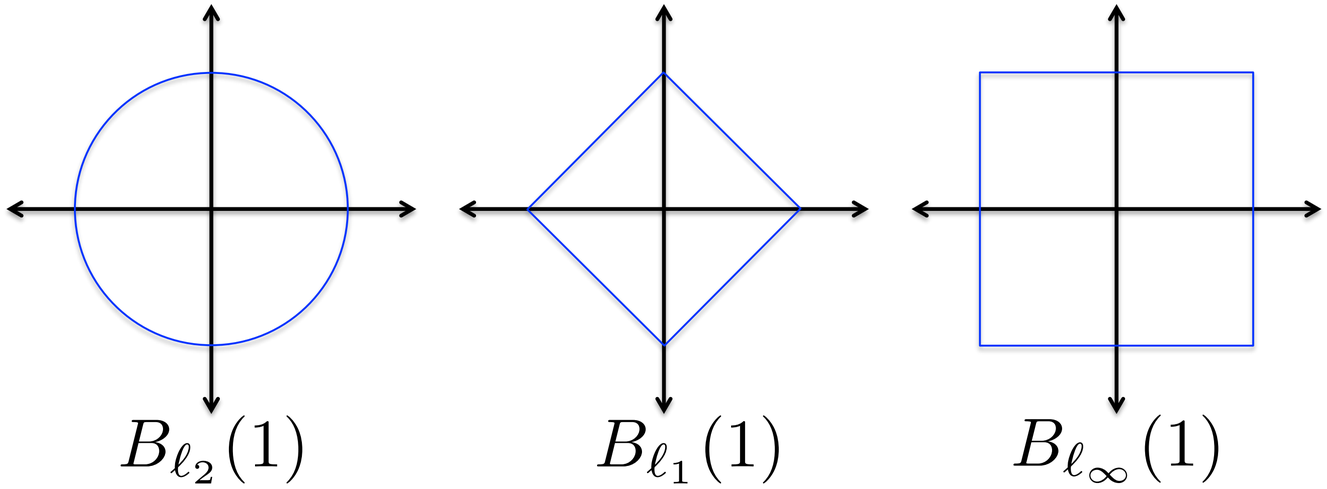
背景知识

Norm范数的定义

ℓ_p-norm balls

Definition (ℓ_p ball)

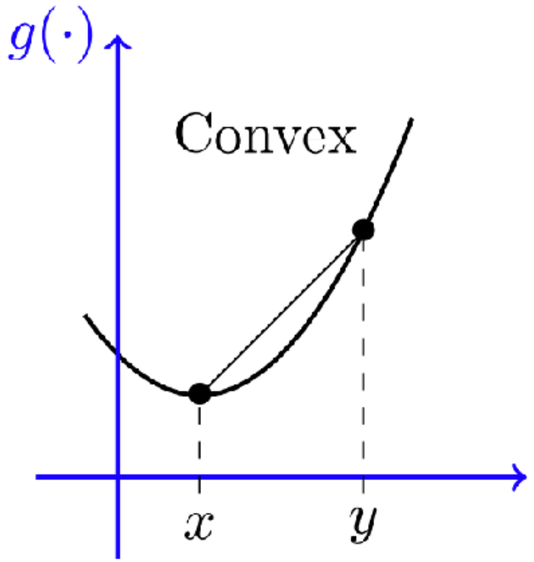
$$\epsilon \geq 0, \quad B_{\ell_p}(\epsilon) = B_p(\epsilon) = \{\mathbf{a} \mid \|\mathbf{a}\|_p \leq \epsilon\}$$



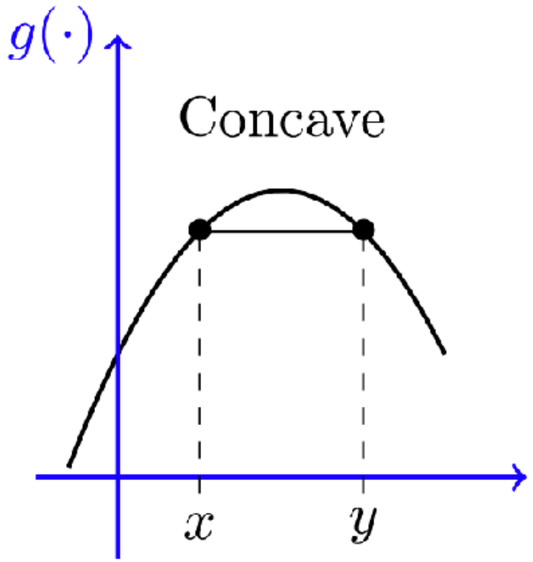
$B_p(1)$ is referred to as the *unit ball* (i.e, $\epsilon = 1$) .

背景知识

凸函数



$$g(\alpha x + (1-\alpha)y) \leq \alpha g(x) + (1-\alpha)g(y).$$



$$g(\alpha x + (1-\alpha)y) \geq \alpha g(x) + (1-\alpha)g(y).$$

背景知识

无约束的优化问题

$$\min \quad f_0(x)$$

梯度下降、牛顿法等

有约束的优化问题

$$\begin{aligned} \min \quad & f_0(x) \\ s.t. \quad & f_i(x) \le 0, \quad i = 1, \dots, m \\ & h_i(x) = 0, \quad i = 1, \dots, p \end{aligned}$$

拉格朗日法

背景知识

有约束的优化问题

$$\begin{aligned} \min \quad & f_0(x) \\ s.t. \quad & f_i(x) \le 0, \quad i = 1, \dots, m \\ & h_i(x) = 0, \quad i = 1, \dots, p \end{aligned}$$

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

$$\max_{\lambda \succeq 0, \nu} \min_{x} L(x, \lambda, \nu)$$

拉格朗日乘数法

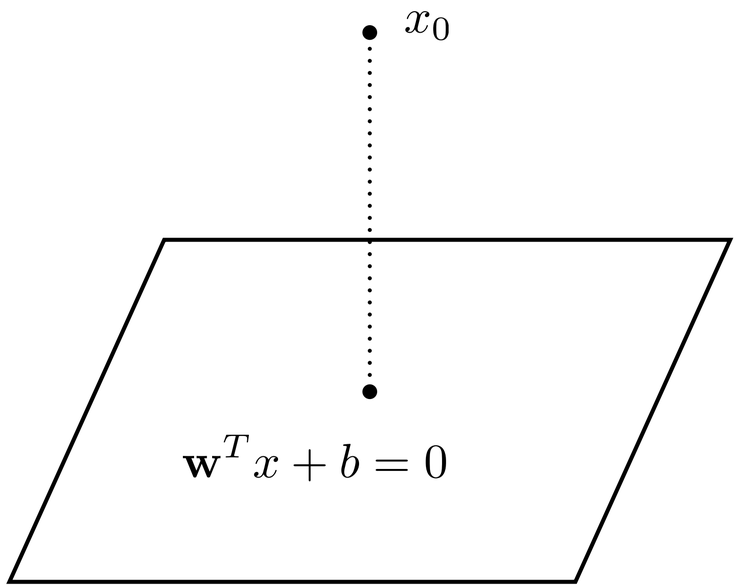
$$\begin{aligned} f_i(x^*) &\le 0, \quad i = 1, \dots, m \\ h_i(x^*) &= 0, \quad i = 1, \dots, p \\ \lambda_i^* &\ge 0, \quad i = 1, \dots, m \\ \lambda_i^* f_i(x^*) &= 0, \quad i = 1, \dots, m \end{aligned}$$

$$\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} \nu_i^* \nabla h_i(x^*) = 0$$

KKT条件

支撑向量机

点到平面/超平面的距离



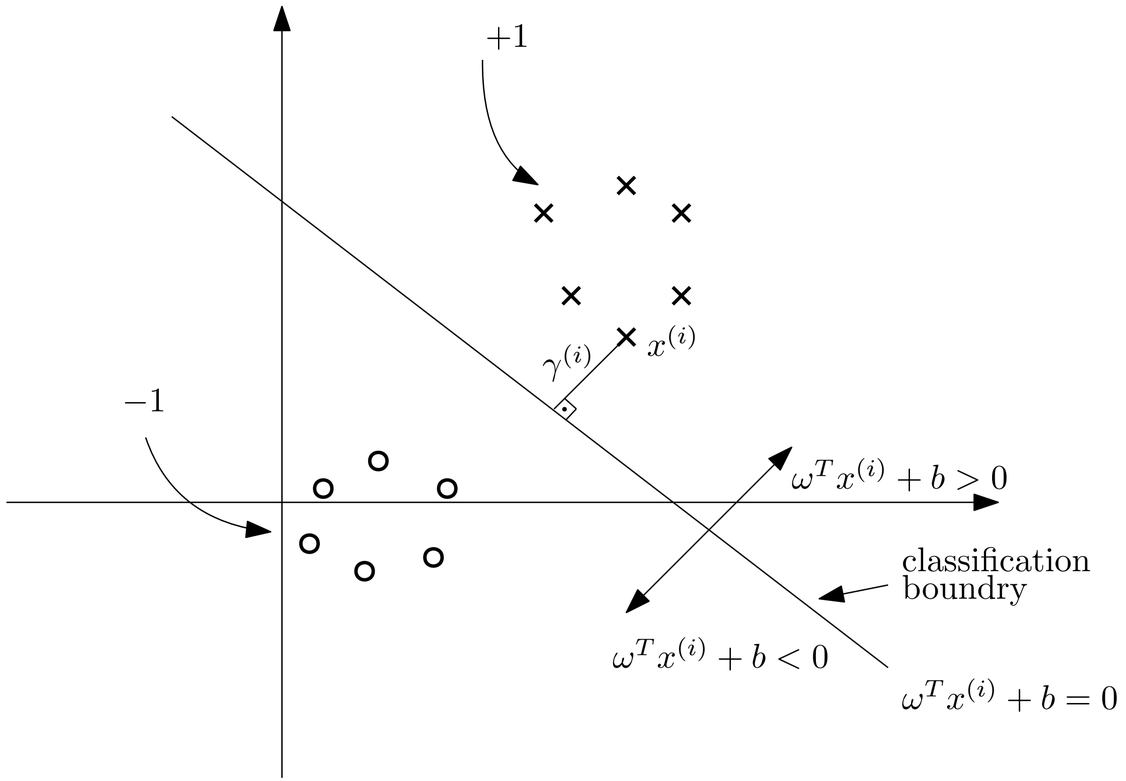
$$\min \frac{1}{2}\|\|\boldsymbol{x} - \boldsymbol{x}_0\|^2$$
$$\text{s.t. } \boldsymbol{w}^T\boldsymbol{x} + b = 0$$

$$\min_{\boldsymbol{x}} \max_{\beta} \frac{1}{2}\|\|\boldsymbol{x} - \boldsymbol{x}_0\|^2 + \beta(\boldsymbol{w}^T\boldsymbol{x} + b)$$

$$\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \beta) = \boldsymbol{x}^* - \boldsymbol{x}_0 + \beta^*\boldsymbol{w} = 0$$
$$\nabla_{\beta} L(\boldsymbol{x}, \beta) = \boldsymbol{w}^T\boldsymbol{x}^* + b = 0$$

$$\gamma_0 = ||\boldsymbol{x}_0 - \boldsymbol{x}^*|| = \left|\frac{\boldsymbol{w}^T\boldsymbol{x}_0 + b}{||\boldsymbol{w}||^2}\boldsymbol{w}\right| = \frac{|\boldsymbol{w}^T\boldsymbol{x}_0 + b|}{||\boldsymbol{w}||}$$

支撑向量机

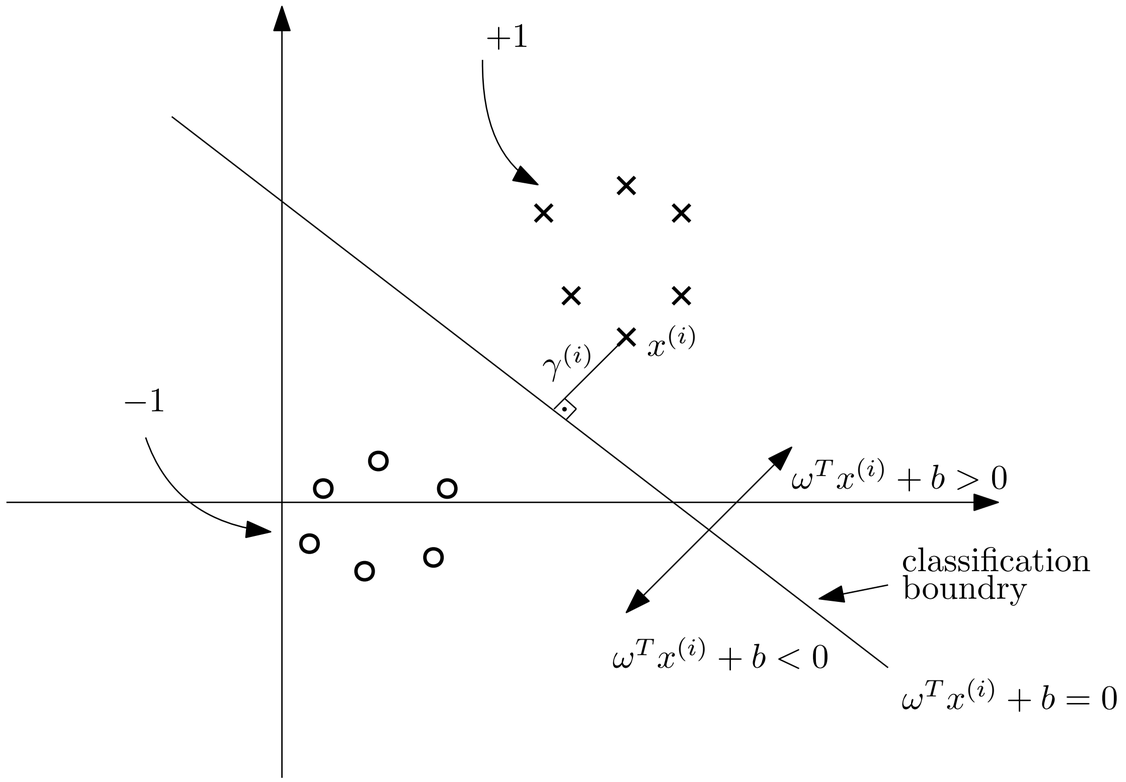


$$\min_{\{\boldsymbol{w},b\}} \quad \frac{1}{2}||\boldsymbol{w}||^2$$

$$\text{s.t.} \quad y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq 1, \quad \forall i$$

损失函数

支撑向量机



$$\begin{aligned} \max_{\{\boldsymbol{w},b\}} \quad & \gamma \\ \text{s.t.} \quad & \gamma^{(i)} \geq \gamma, \quad \forall i \\ & y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq 0, \quad \forall i \end{aligned}$$

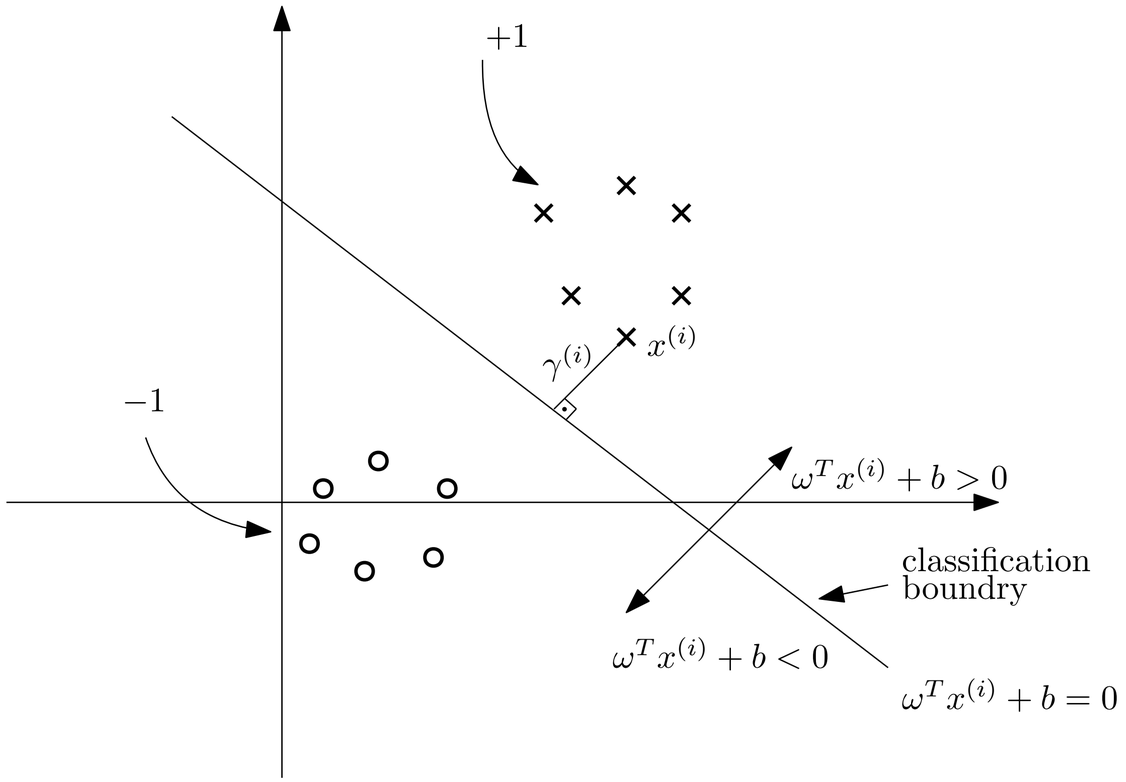
1、让所有点到平面的距离都比γ要大

2、第二个条件分类正确

$$\gamma_0 = ||\boldsymbol{x}_0 - \boldsymbol{x}^*|| = \left|\frac{\boldsymbol{w}^T\boldsymbol{x}_0 + b}{||\boldsymbol{w}||^2}\boldsymbol{w}\right| = \frac{|\boldsymbol{w}^T\boldsymbol{x}_0 + b|}{||\boldsymbol{w}||}$$

点到平面的距离

支撑向量机



代入点到平面的距离

$$\max_{\{\boldsymbol{w},b\}} \quad \gamma$$

$$\text{s.t.} \quad \frac{|\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b|}{||w||} \geq \gamma, \quad \forall i$$

$$y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq 0, \quad \forall i$$

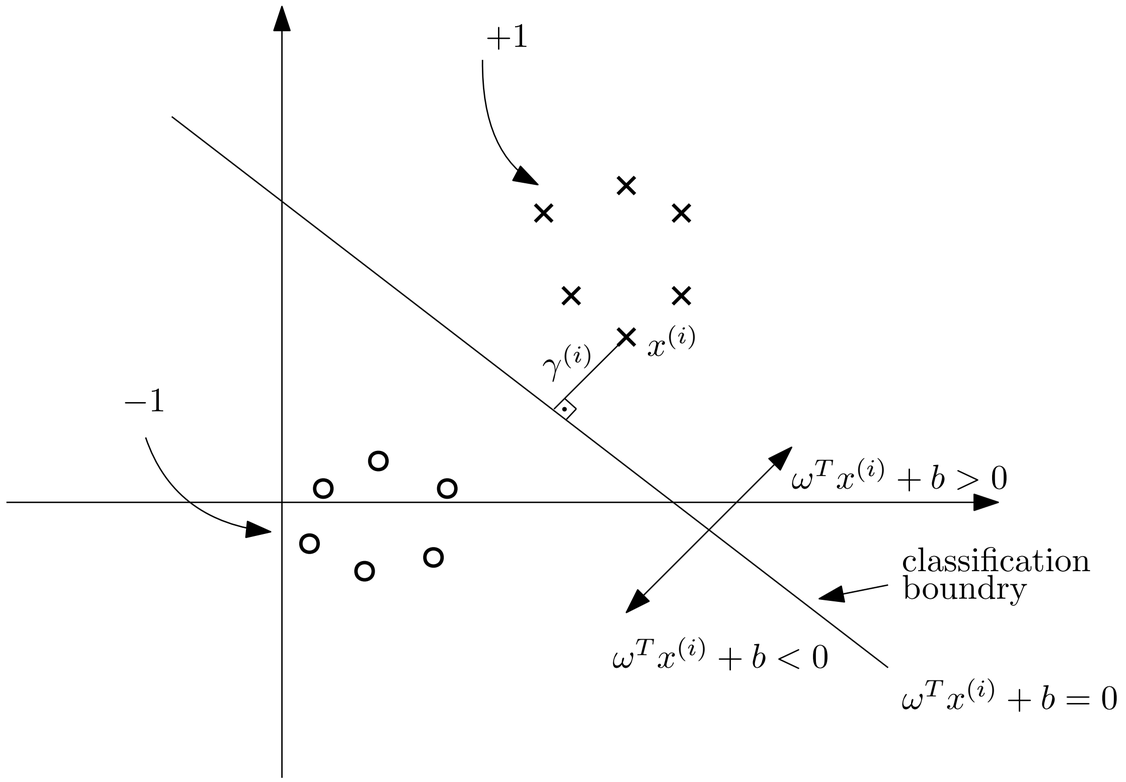
$$y^{(i)}\left(\boldsymbol{w}^{*T}\boldsymbol{x}^{(i)} + b^*\right) = |\boldsymbol{w}^{*T}\boldsymbol{x}^{(i)} + b^*|$$

对于分对的样本

$$\max_{\{\boldsymbol{w},b,\gamma\}} \quad \gamma$$

$$\text{s.t.} \quad y^{(i)}\left(\boldsymbol{w}^{*T}\boldsymbol{x}^{(i)} + b^*\right) \geq \gamma||\boldsymbol{w}||$$

支撑向量机



令 $\gamma' = \gamma||\boldsymbol{w}||$

$$\max_{\{\boldsymbol{w},b,,\gamma'\}} \quad \frac{\gamma'}{||\boldsymbol{w}||}$$
$$\text{s.t.} \quad y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq \gamma'$$

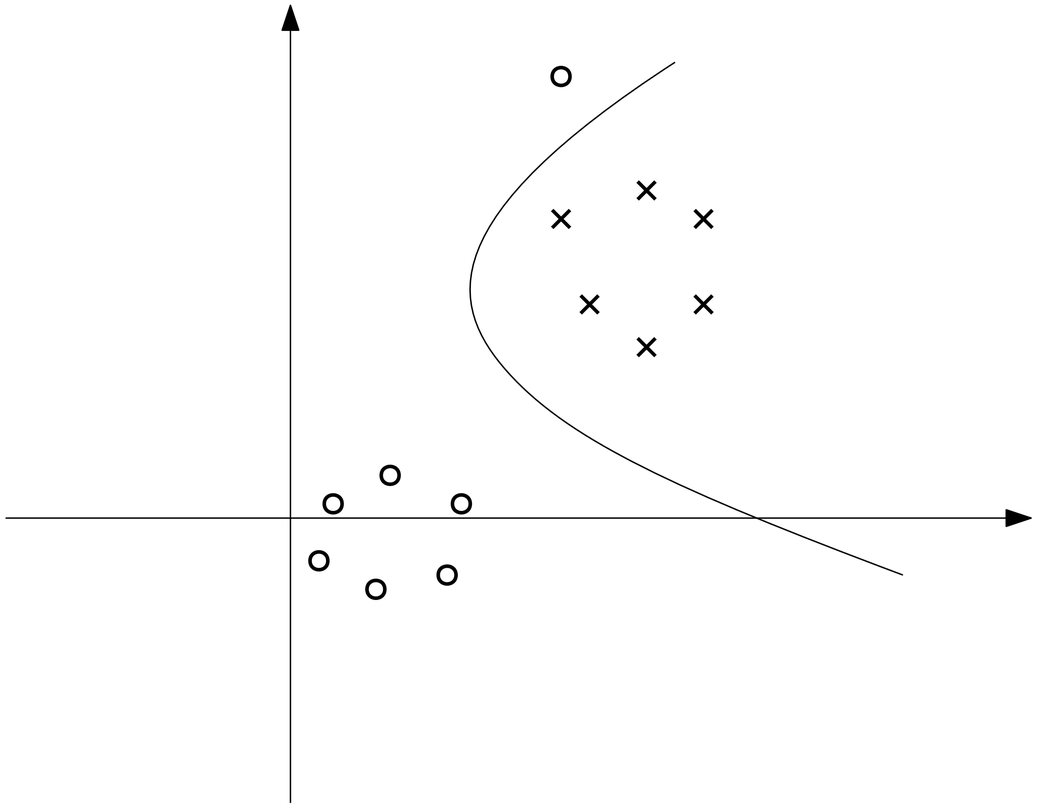
对间隔进行归一

$$\max_{\{\boldsymbol{w},b\}} \quad \frac{1}{||\boldsymbol{w}||}$$
$$\text{s.t.} \quad y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq 1, \quad \forall i$$

从max到min

$$\min_{\{\boldsymbol{w},b\}} \quad \frac{1}{2}||\boldsymbol{w}||^2$$
$$\text{s.t.} \quad y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq 1, \quad \forall i$$

支撑向量机



Soft SVM

$$\min_{\{\boldsymbol{w},b\}} \quad \frac{1}{2}||\boldsymbol{w}||^2 + c\sum_{i=1}^{m}\xi_i$$

$$\text{s.t.} \quad y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq 1 - \xi_i, \quad \forall i$$

$$\xi_i \geq 0, \quad \forall i$$

支撑向量机—核函数

Hard SVM

$$\min_{\{\boldsymbol{w},b\}} \quad \frac{1}{2}||\boldsymbol{w}||^2$$
$$\text{s.t.} \quad y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)}+b\right)\geq 1, \quad \forall i$$

$$\min_{\{\boldsymbol{w},b\}}\max_{\alpha_i\geq 0} \quad \frac{1}{2}||\boldsymbol{w}||^2+\sum_{i=1}^{m}\alpha_i\left(1-y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)}+b\right)\right)$$
$$\max_{\alpha_i\geq 0}\min_{\{\boldsymbol{w},b\}} \quad \frac{1}{2}||\boldsymbol{w}||^2+\sum_{i=1}^{m}\alpha_i\left(1-y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)}+b\right)\right)$$

$$\nabla_{\boldsymbol{w}}\left(\frac{1}{2}||\boldsymbol{w}||^2+\sum_{i=1}^{m}\alpha_i\left(1-y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)}+b\right)\right)\right)=\boldsymbol{w}-\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)}=0$$
$$\rightarrow \boldsymbol{w}^*=\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)}$$
$$\nabla_b\left(\frac{1}{2}||\boldsymbol{w}||^2+\sum_{i=1}^{m}\alpha_i\left(1-y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)}+b\right)\right)\right)=-\sum_{i=1}^{m}\alpha_i y^{(i)}=0$$

支撑向量机—核函数

Hard SVM

$$\max_{\alpha_i \geq 0} \min_{\{\boldsymbol{w},b\}} \quad \frac{1}{2}||\boldsymbol{w}||^2 + \sum_{i=1}^{m} \alpha_i \left(1 - y^{(i)} \left(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b\right)\right)$$

$$\boldsymbol{w}^* = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$$

$$-\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

$$\max_{\alpha_i \geq 0} \quad \frac{1}{2}||\sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}||^2 + \sum_{i=1}^{m} \alpha_i - \sum_{i=1}^{m} \alpha_i y^{(i)} (x^{(i)})^T \left(\sum_{j=1}^{m} \alpha_j y^{(j)} x^{(j)}\right) - \sum_{i=1}^{m} \alpha_i y^{(i)} b^*$$

$$||a_1 + a_2||^2 = ||a_1||^2 + ||a_2||^2 + 2a_1^T a_2 = \sum_{i,j=1}^{2} a_i^T a_j$$

$$\max_{\alpha_i \geq 0} \quad \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)} + \sum_{i=1}^{m} \alpha_i - \sum_{i,j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)}$$

$$\max_{\alpha_i \geq 0} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} < x^{(i)}, x^{(j)} >$$

$$\text{s.t.} \quad \sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

支撑向量机—核函数

Hard SVM

对新样本的分类

$$\boldsymbol{w}^* = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$$

$$(\boldsymbol{w}^*)^T x^{(new)} + b^* = (\sum_{i=1}^{m} \alpha^* y^{(i)} x^{(i)})^T x^{(new)} + b^*$$

$$= \sum_{i=1}^{m} \alpha^* y^{(i)} < x^{(i)}, x^{(new)} > + b^*$$

支撑向量机—核函数

对新样本的分类

$$\sum_{i=1}^{m} \alpha^* y^{(i)} < x^{(i)}, x^{(new)} > + b^*$$

$$X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$\phi(x) = \begin{bmatrix} x_1^2 \\ x_1 x_2 \\ x_1 x_3 \\ x_2 x_1 \\ x_2 x_2 \\ x_2 x_3 \\ x_3 x_1 \\ x_3 x_2 \\ x_3 x_3 \end{bmatrix}$$

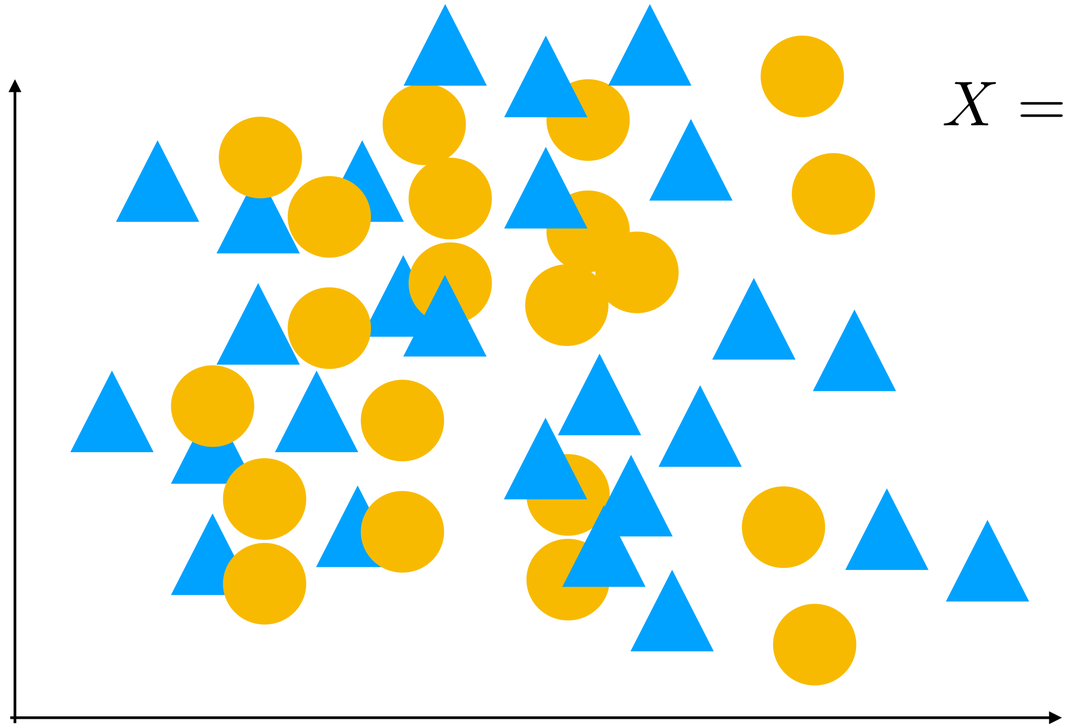
$$\max_{\alpha_i \geq 0} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} < x^{(i)}, x^{(j)} >$$

$$\text{s.t.} \quad \sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

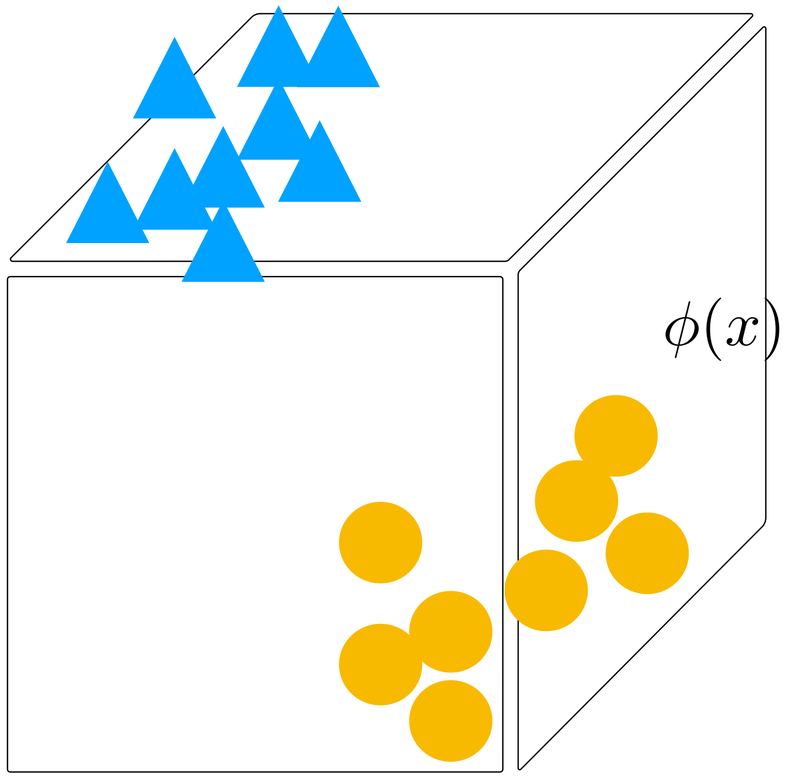
$$\max \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} < \phi(x^{(i)}), \phi(x^{(j)}) >$$

支撑向量机—核函数

特征的升维度



$$X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$



$$\phi(x) = \begin{bmatrix} x_1^2 \\ x_1x_2 \\ x_1x_3 \\ x_2x_1 \\ x_2x_2 \\ x_2x_3 \\ x_3x_1 \\ x_3x_2 \\ x_3x_3 \end{bmatrix}$$

模型的选择与评估

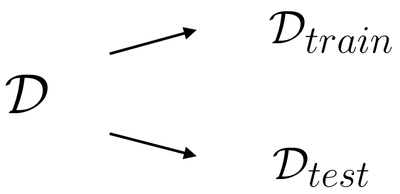
数据集处理

训练集(training set)

如何评估predictor的好坏?

构造测试集(testing set)，用测试集中的样本(test sample)来检验predictor的好坏。测试集应与训练集互斥，测试样本不会用于训练过程。

通常从给定数据集(dataset)中划分出训练集和测试集。



留出法 (Hold-out)

将数据集**若干次随机**划分为**无交集**的训练集和测试集，重复实验评估后取平均值作为结果。

通常训练集占总样本的80%，测试集占总样本的20%。

Hold-out 受到数据划分方式的影响，模型容易偏向于训练集中出现次数多的样本。

交叉验证法 (k-fold cross-validation)

将数据集随机等分为k-fold，每次取一个fold作为测试集，输出k个结果的平均值。

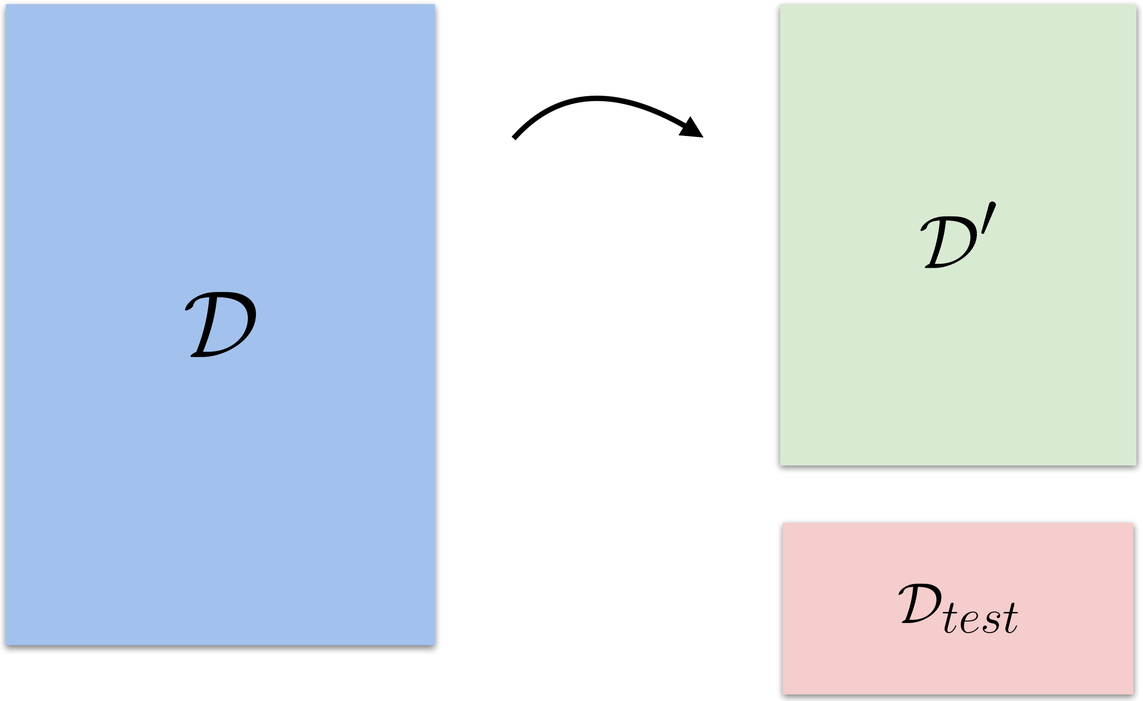
留一法 (Leave-One-Out)

k-fold cross validation的特殊形式（k等于样本总数），即每次取出一个样本作为测试样本，剩余样本作为训练集。

LOO不受数据划分方式的影响，与原始数据集分布近似，其评估结果通常认为比较准确，但当数据集较大时，计算开销也随之增大。

自助法 (Bootstrapping)

有放回地重复抽样m次



原数据集D中部分样本重复在D'中出现，部分样本从未在D'中出现。

原数据集D中不被D'包含的数据作为测试集

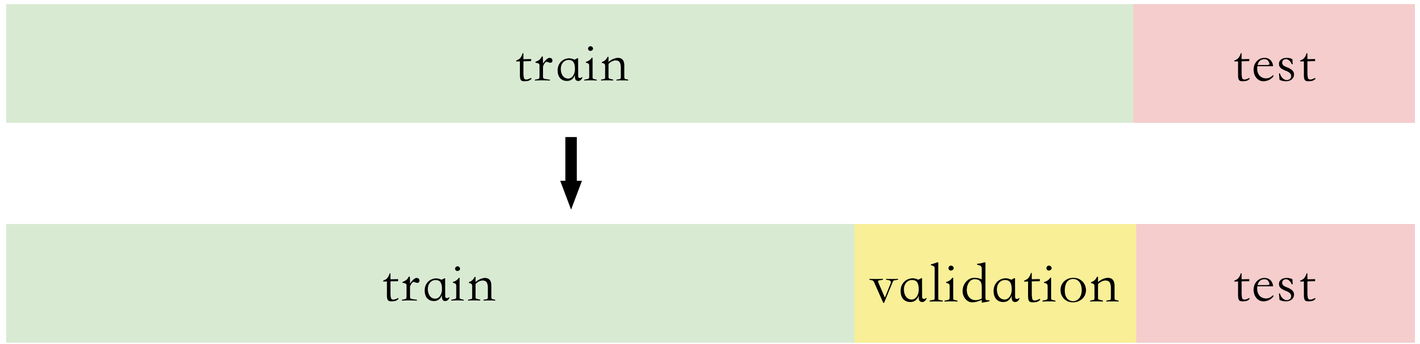
样本不被取到的概率为 $\left(1-\frac{1}{m}\right)^m$

$$\lim_{m\to\infty}\left(1-\frac{1}{m}\right)^m \to \frac{1}{e} \approx 0.368$$

自助法在数据集较小难以划分时比较有效，但改变了原有数据集的分布，会引入估计误差。

验证集(validation set)

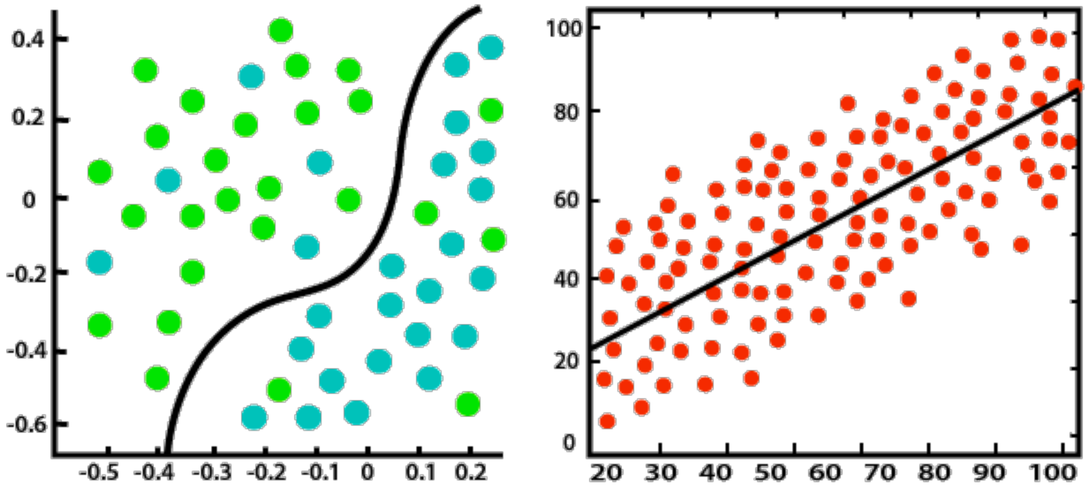
模型评估与选择中，常常从训练集中拆分一部分数据作为验证集(validation set)。



验证集上进行调参和模型的选择，而测试集是实际应用中的数据，用于最终的性能度量。

性能度量

为了衡量一个机器学习模型的好坏，需要模型对**测试集**中的每一个样本进行预测，并根据预测结果计算评价分数，不同学习任务模型性能的评价标准不同。



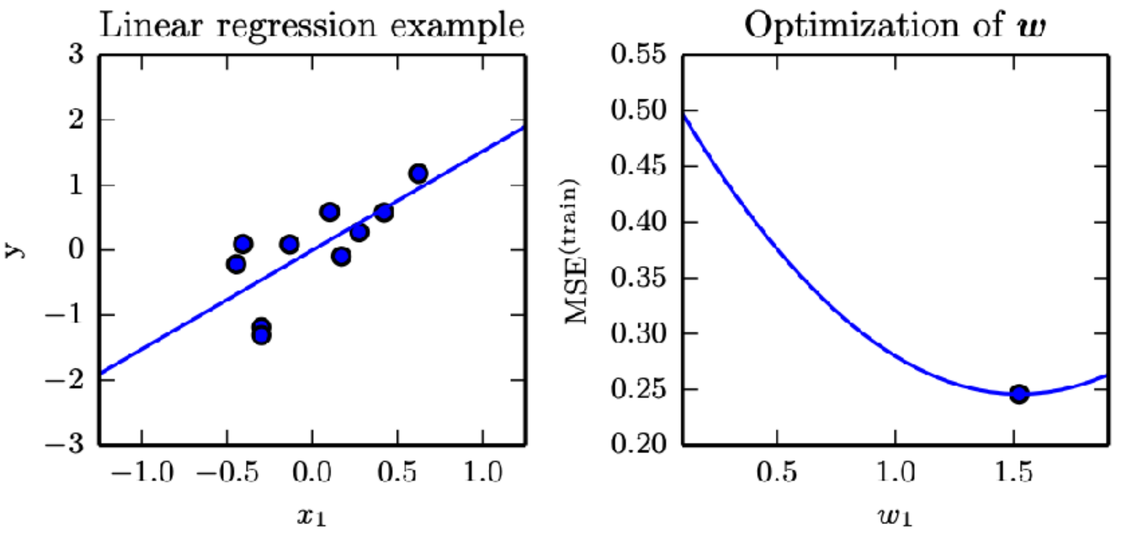
回归模型

比较学习器预测结果$f(x)$与真实标记y

均方误差(mean squared error, MSE)

$$E(f;D)=\frac{1}{N}\sum_{n=1}^{N}(f(x_n)-y_n)^2$$

$$E(f;D)\int_{x\sim D}(f(x)-y)^2p(x)dx$$



其中D是数据分布，$p(x)$是概率密度函数。

分类模型

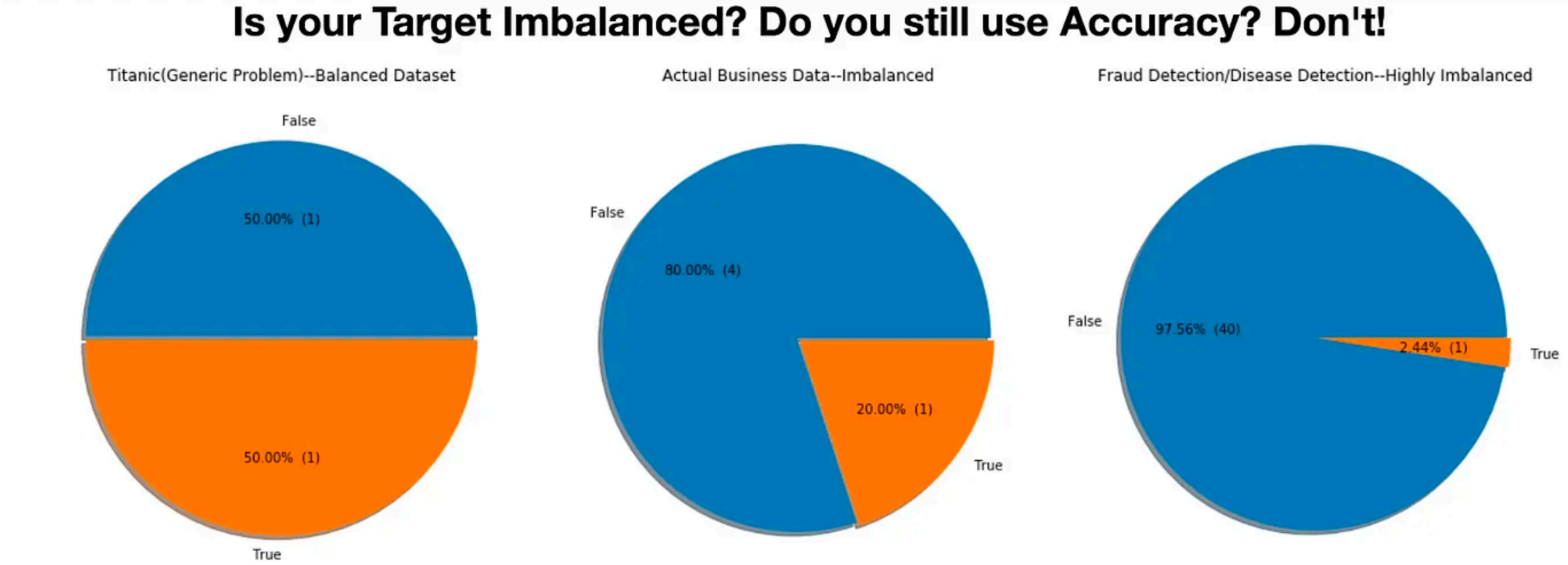
错误率(error rate)

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{M}\mathbb{I}\left(f(x_i)\neq y_i\right)$$

准确率(accuracy)

$$acc(f;D)=\frac{1}{m}\sum_{i=1}^{M}\mathbb{I}\left(f(x_i)=y_i\right)=1-E(f;D)$$

分类模型


Is your Target Imbalanced? Do you still use Accuracy? Don't!
Titanic(Generic Problem)--Balanced Dataset
False
50.00% (1)
50.00% (1)
True
Actual Business Data--Imbalanced
False
80.00% (4)
20.00% (1)
True
Fraud Detection/Disease Detection--Highly Imbalanced
False
97.56% (40)
2.44% (1)
True

分类模型

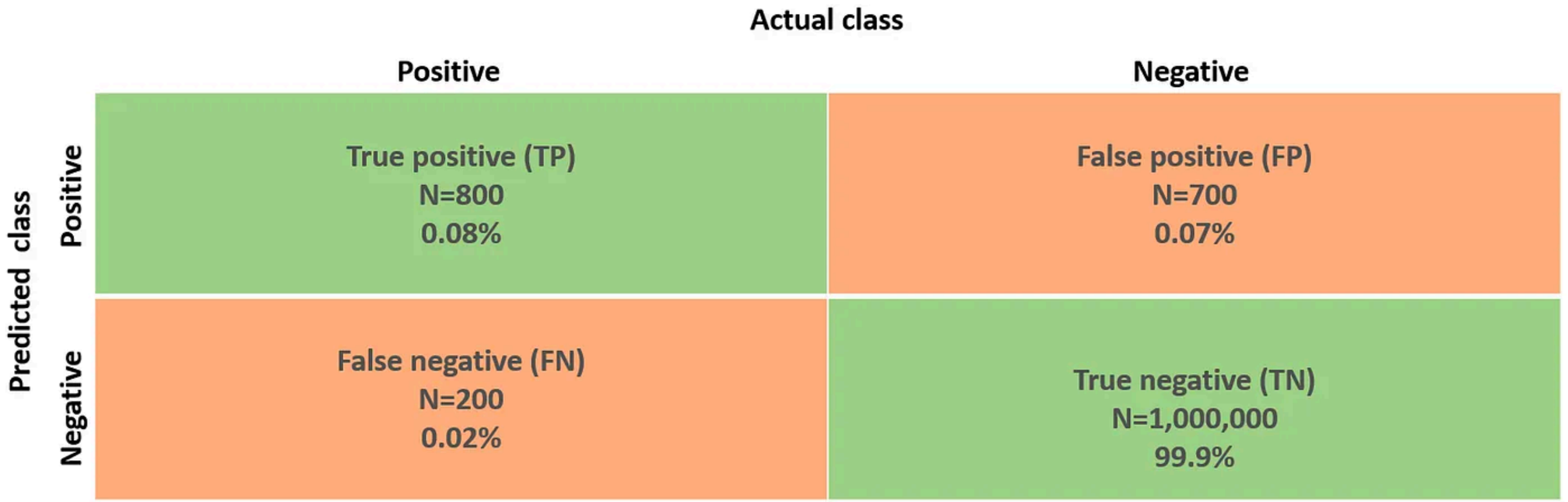
银行的交易记录中仅有0.1%的交易记录是诈骗行为，99.9%是正常交易。为银行开发的信用卡诈骗检测模型进行分类时，可以达到99.9%的判别准确率。然而仅仅考虑模型的判别准确率并不能处理潜在的诈骗行为。

希望考虑的是:

模型判别为诈骗的交易记录中有多少条是真实的诈骗?

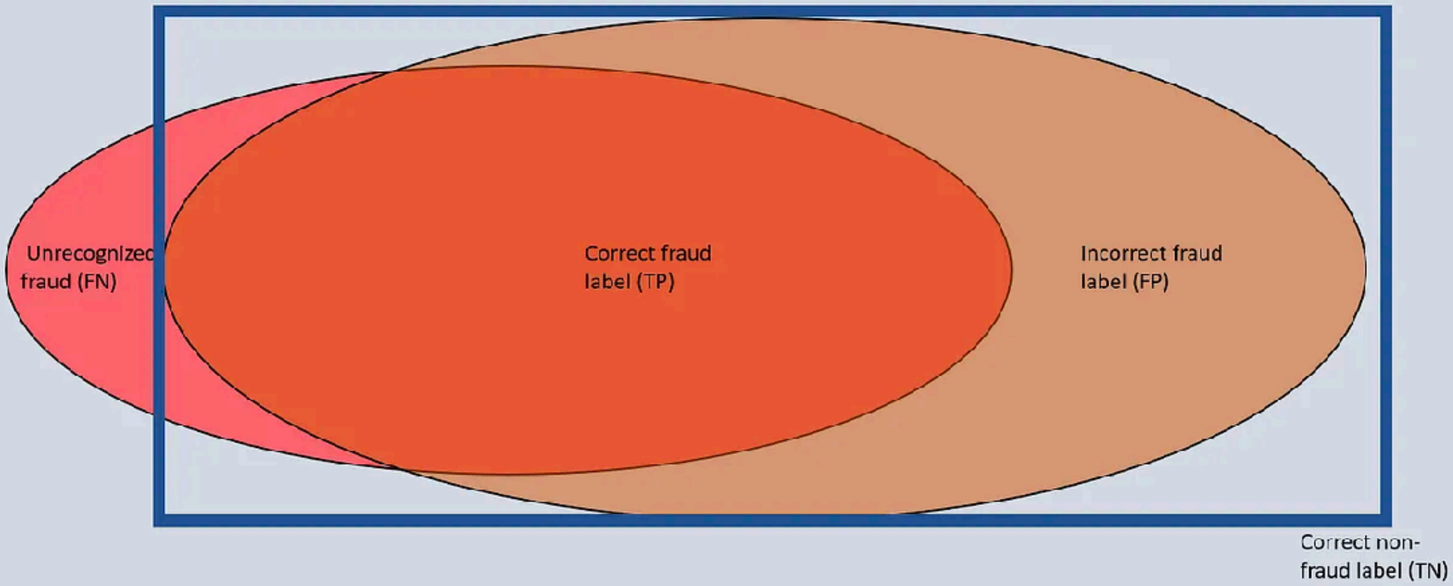
模型判别为正常的交易记录中有多少条反而为诈骗行为？…

混淆矩阵(confusion matrix)



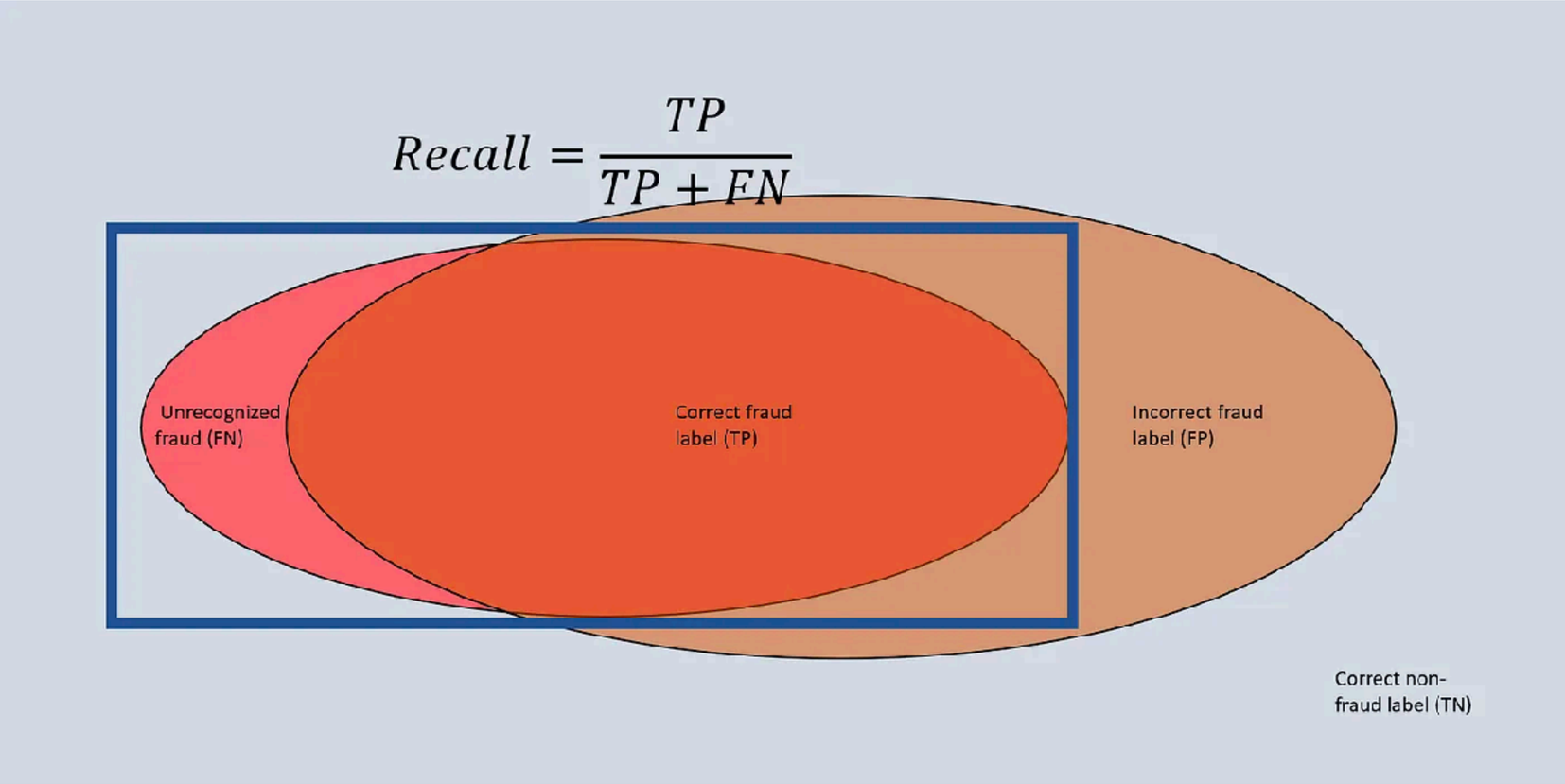
精确率(precision)

$$Precision = \frac{TP}{TP + FP}$$



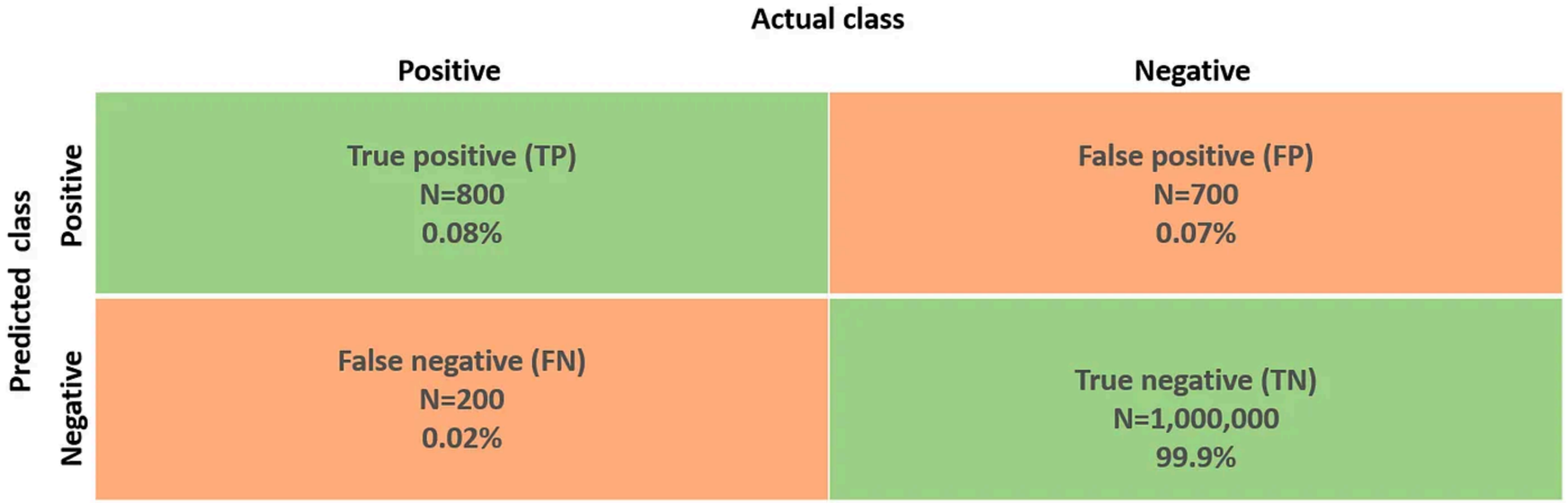
精确率(precision)说明了模型所有判别为诈骗的交易记录中，有多少比例的记录确实为诈骗行为。

召回率(recall)



召回率(recall)说明了所有诈骗记录中，被模型侦测出的诈骗记录的占比。

混淆矩阵



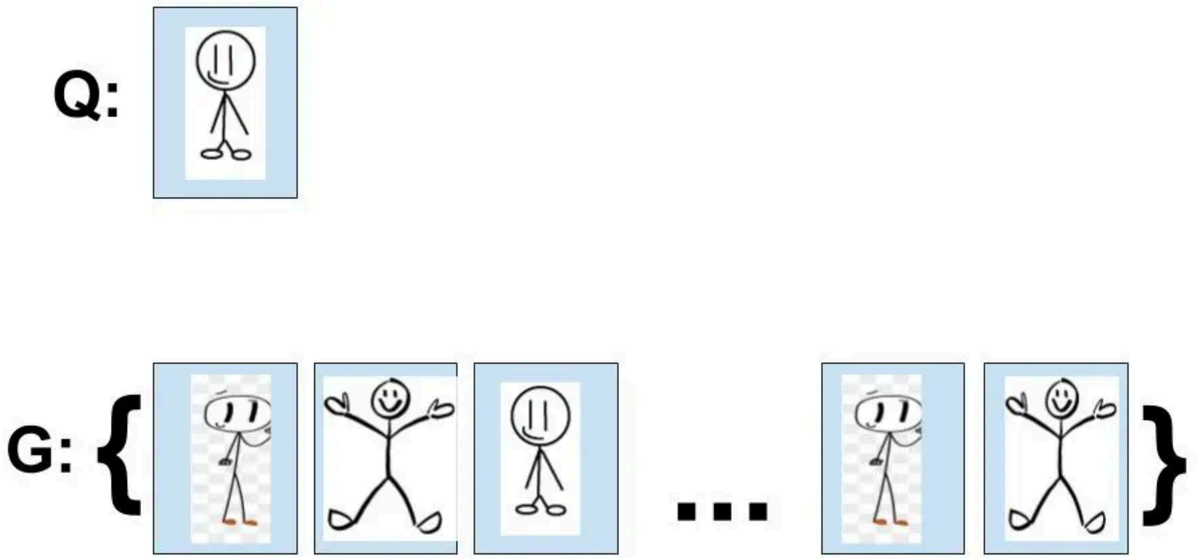
$$precision = \frac{TP}{TP + FP} = \frac{800}{800 + 700} = 0.53$$

$$recall = \frac{TP}{TP + FN} = \frac{800}{800 + 200} = 0.8$$

$$acc = \frac{TP + TN}{TP + TN + FP + FN} = \frac{800 + 1000000}{1001700} = 0.999$$

Precision-Recall

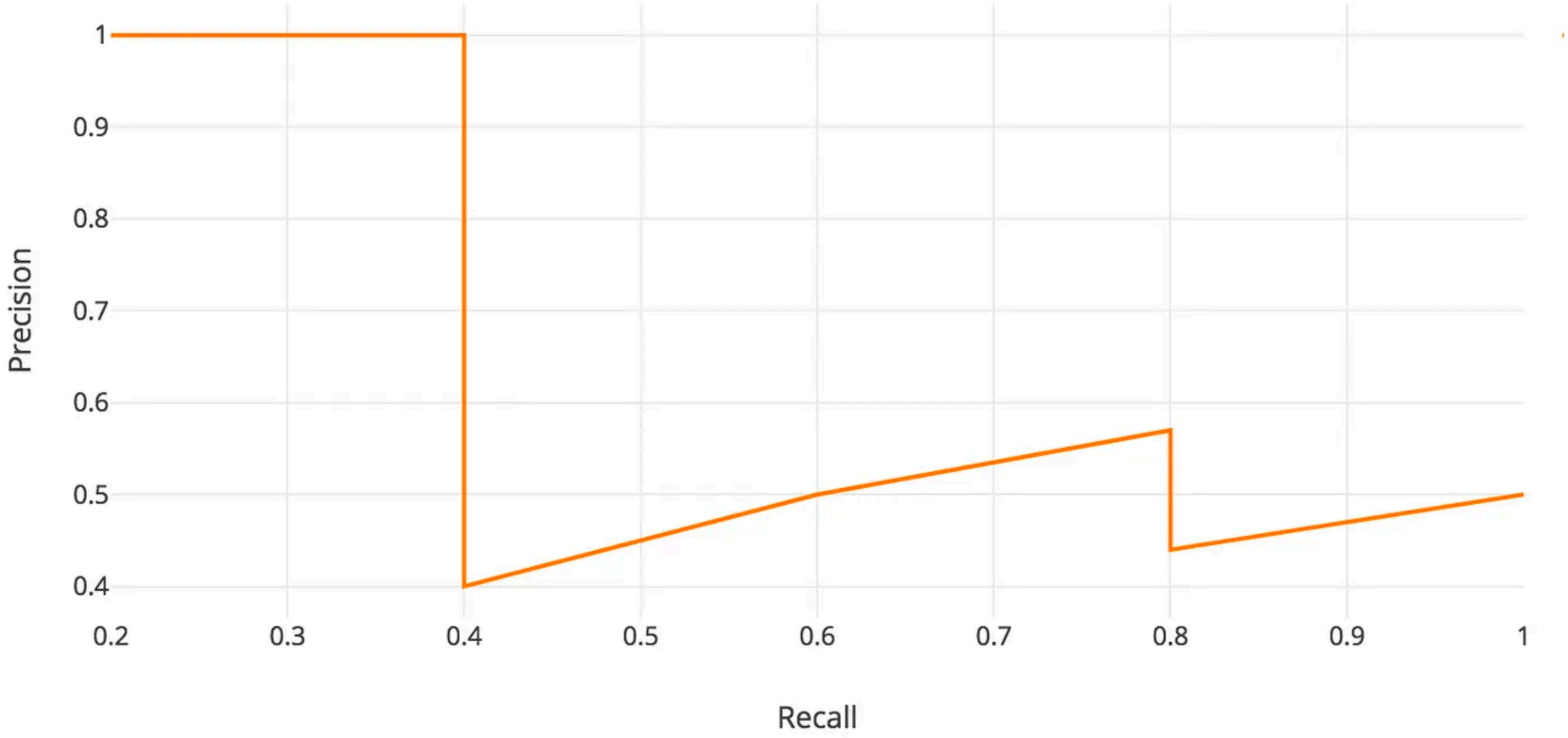
在检索系统中提出



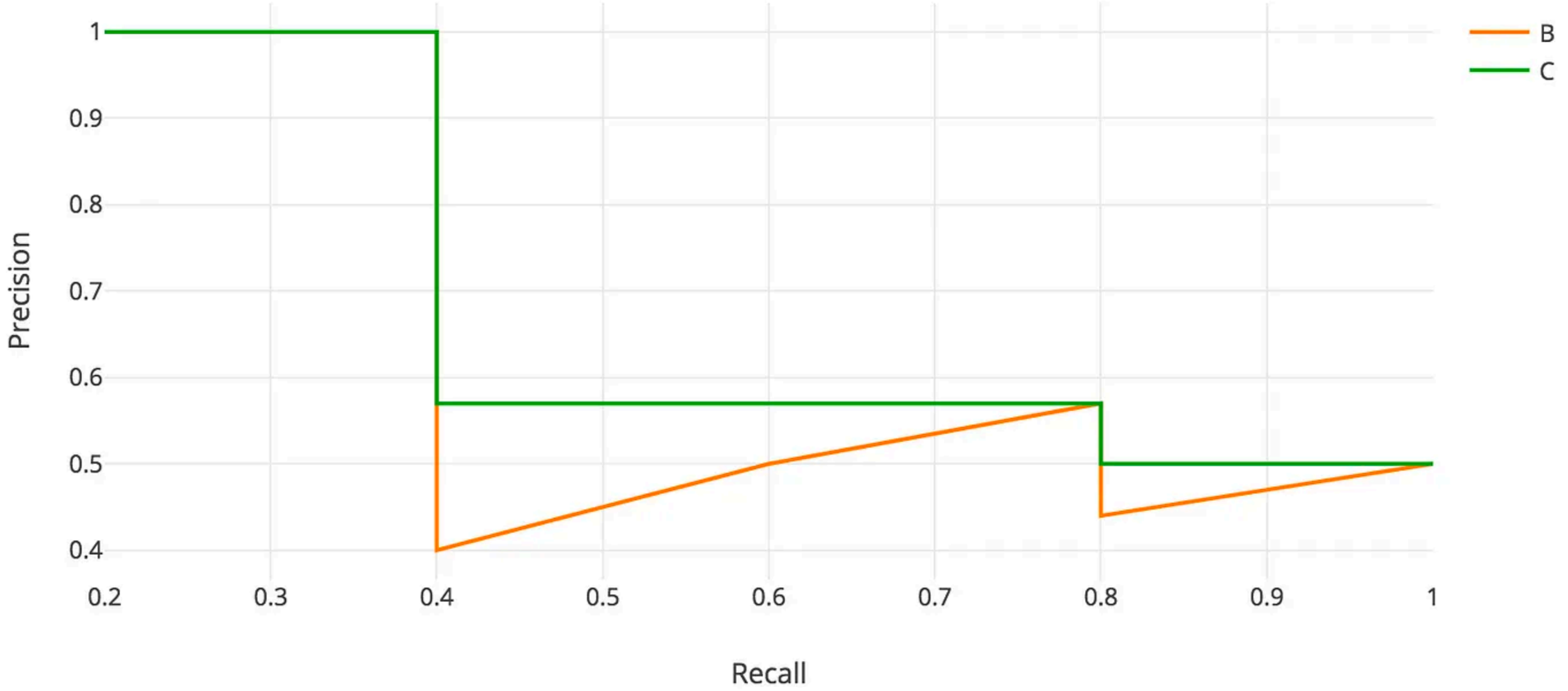
Precision-Recall

排序	是否分对	精度	召回
1	是	1	0.2
2	是	1	0.4
3	否	0.67	0.4
4	否	0.5	0.4
5	否	0.4	0.4
6	是	0.5	0.6
7	是	0.57	0.8
8	否	0.5	0.8
9	否	0.55	0.8
10	是	0.5	1.0

Precision-Recall



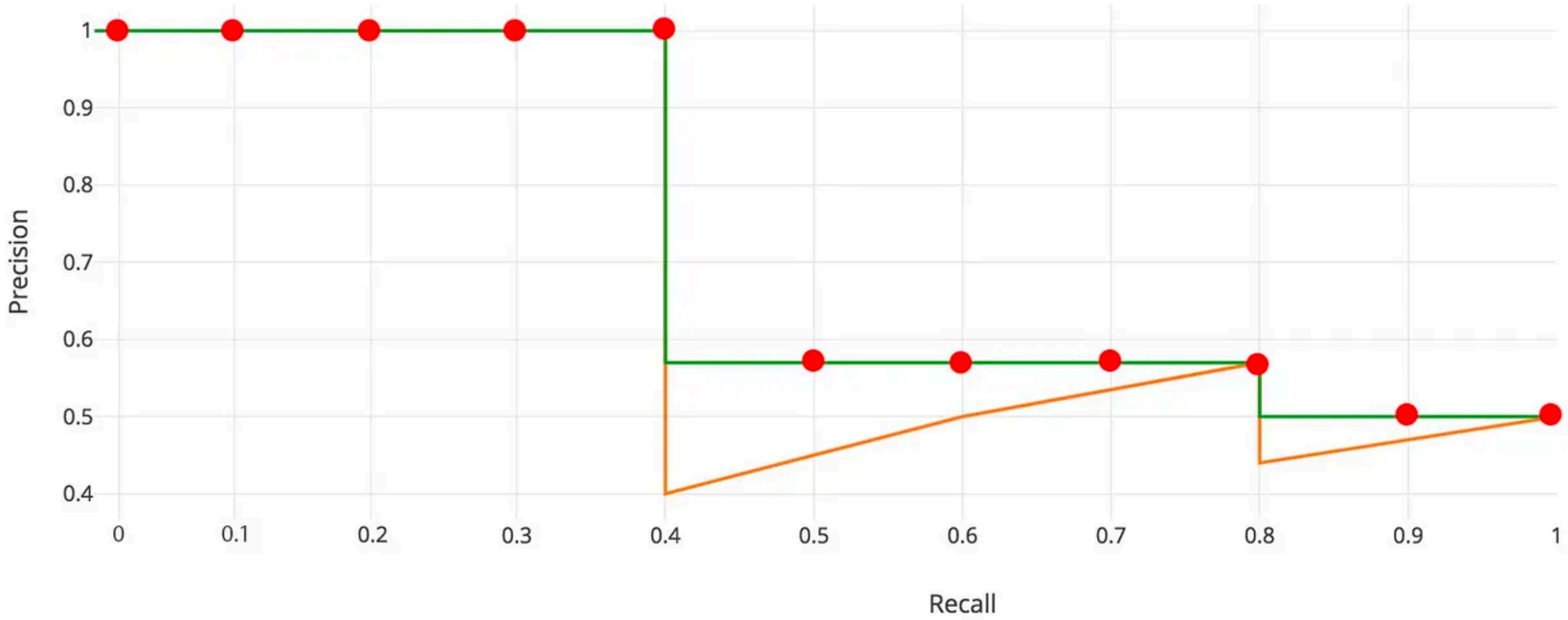
Precision-Recall



召回率 r̂的精度用 ≥ r̂召回下最大的精度

$$p_{interp}(r) = max\ p(\hat{r})\ for\ \hat{r} \geq r$$

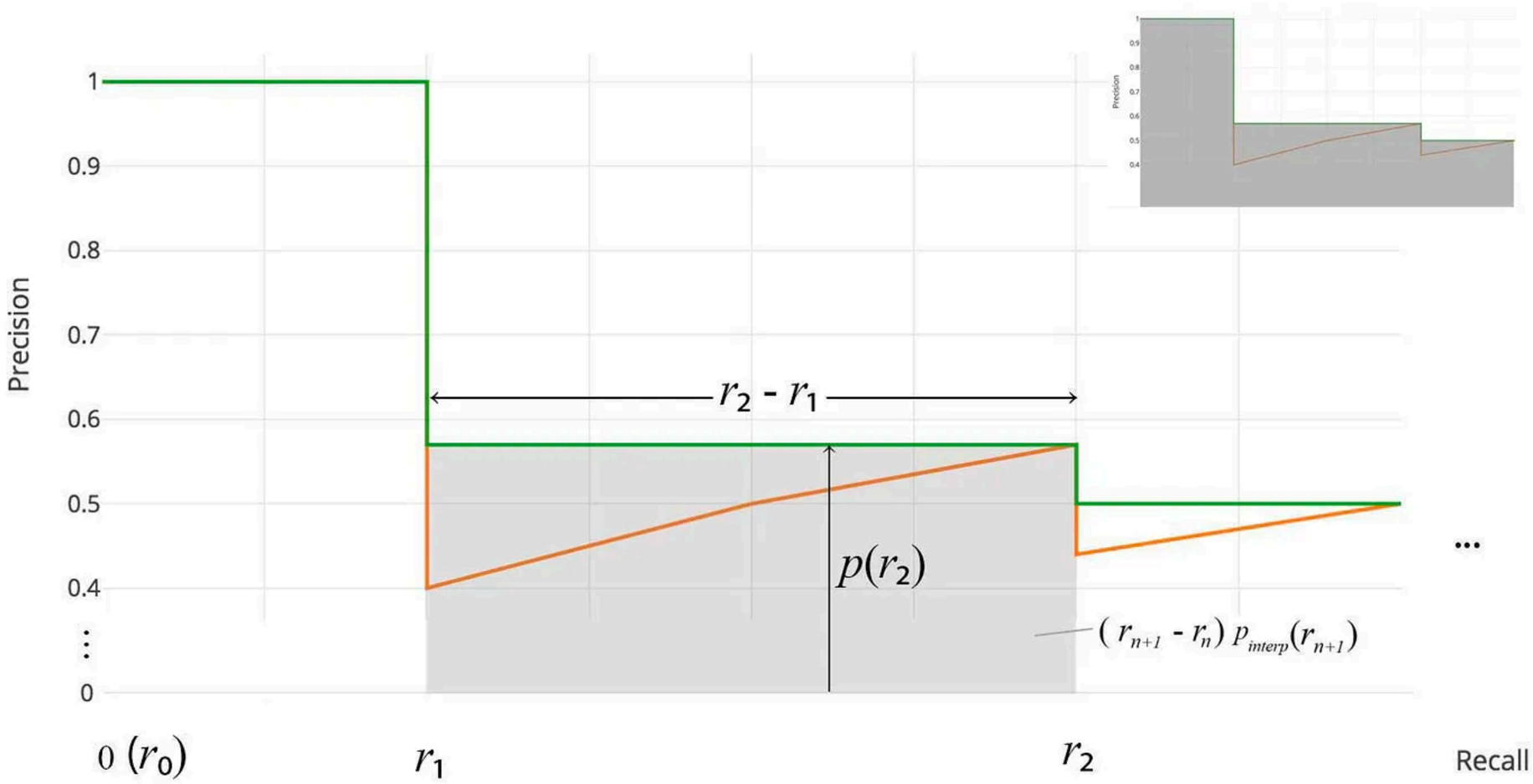
Precision-Recall



$$AP11 = 1/11\ (Ap(0) + Ap(0.1) + \cdots + Ap(1.0))$$

$$AP = (5 \times 1.0 + 4 \times 0.57 + 2 \times 0.5)/11$$

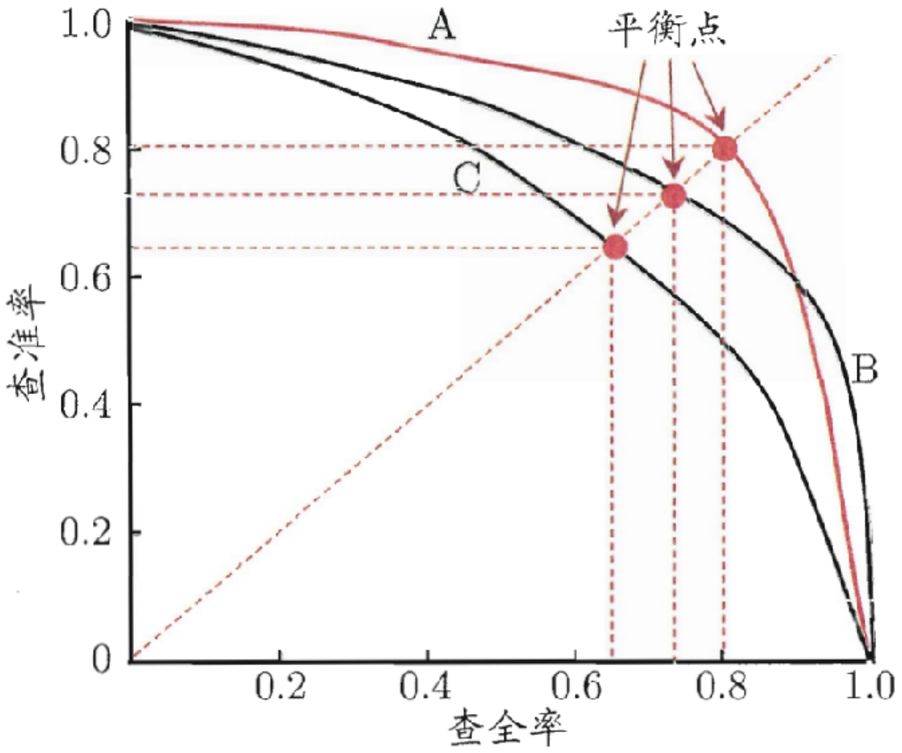
Precision-Recall



$$AP = \sum (r_{n+1} - r_n) p_{interp}(r_{n+1})$$

$$p_{interp}(r_{n+1}) = max\ p(\hat{r})\ for\ \hat{r} \geq r_{n+1}$$

平衡点(Break-Even Point, BEP)



trade-off

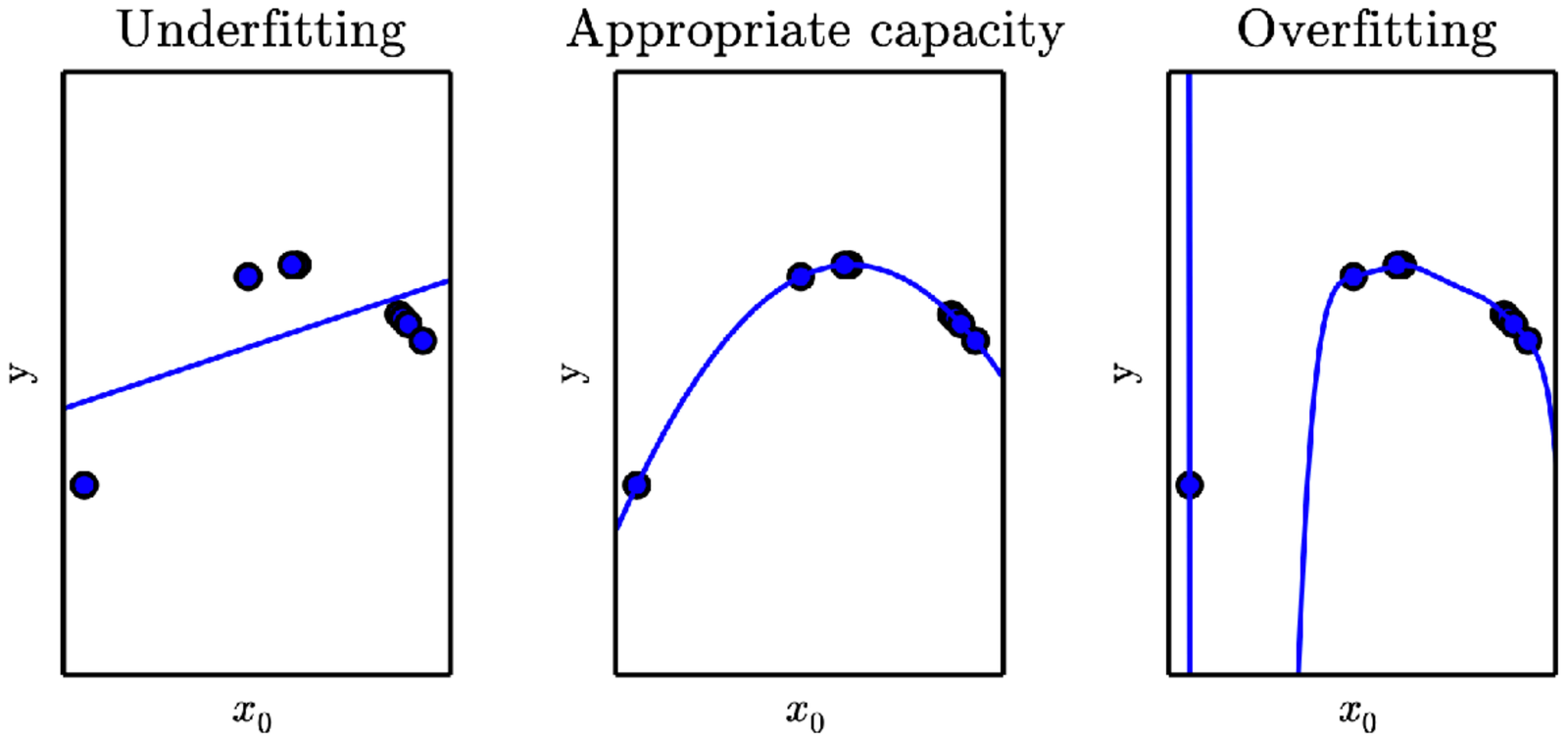
precision ⟷ recall

平衡点(BEP)是P=R时，PR图的取值。

模型选择

奥卡姆剃刀

在同样能够解释已知观测现象的假设中，我们应该挑选“最简单”的那一个



$$y = a_0 + a_1 x \qquad y = a_0 + a_1 x + a_2 x^2 \qquad y = \sum_{i}^{9} a_i x^i$$

误差

经验误差（经验误差）：训练集上的误差 $\frac{1}{m^{(train)}}\|X^{(train)}\ w - y^{(train)}\|$

泛化误差（泛化误差）：测试集上的误差 $\frac{1}{m^{(test)}}\|X^{(test)}\ w - y^{(train)}\|$

泛化：在先前未观测到的输入上表现良好的能力

泛化能力

泛化能力（generalizability）

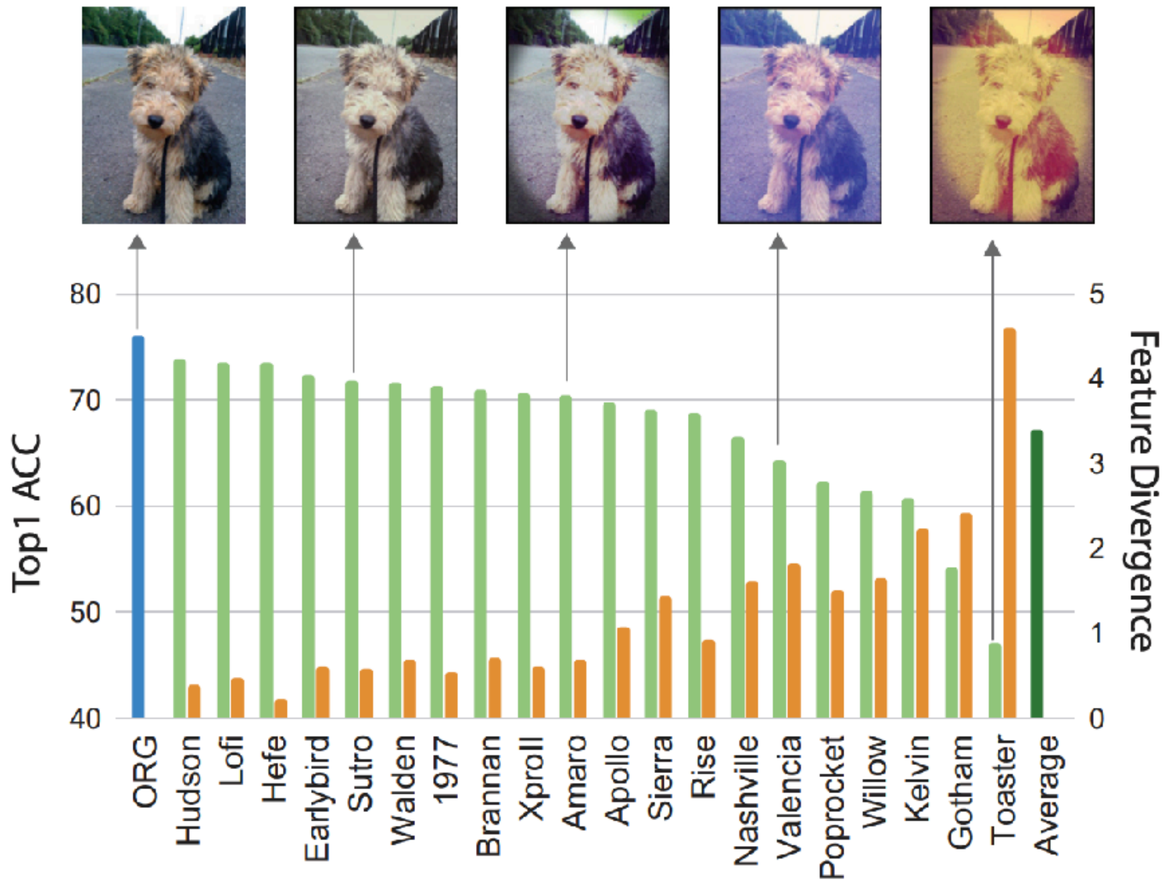
- 指由该方法学习到的模型对未知数据的预测能力
- 是学习方法本质上最重要的性质

泛化误差与泛化能力

- 泛化误差就是泛化能力的一种度量
- 泛化误差越小，则泛化能力越强，学习方法越有效
- 泛化误差是所学习到的模型的期望风险

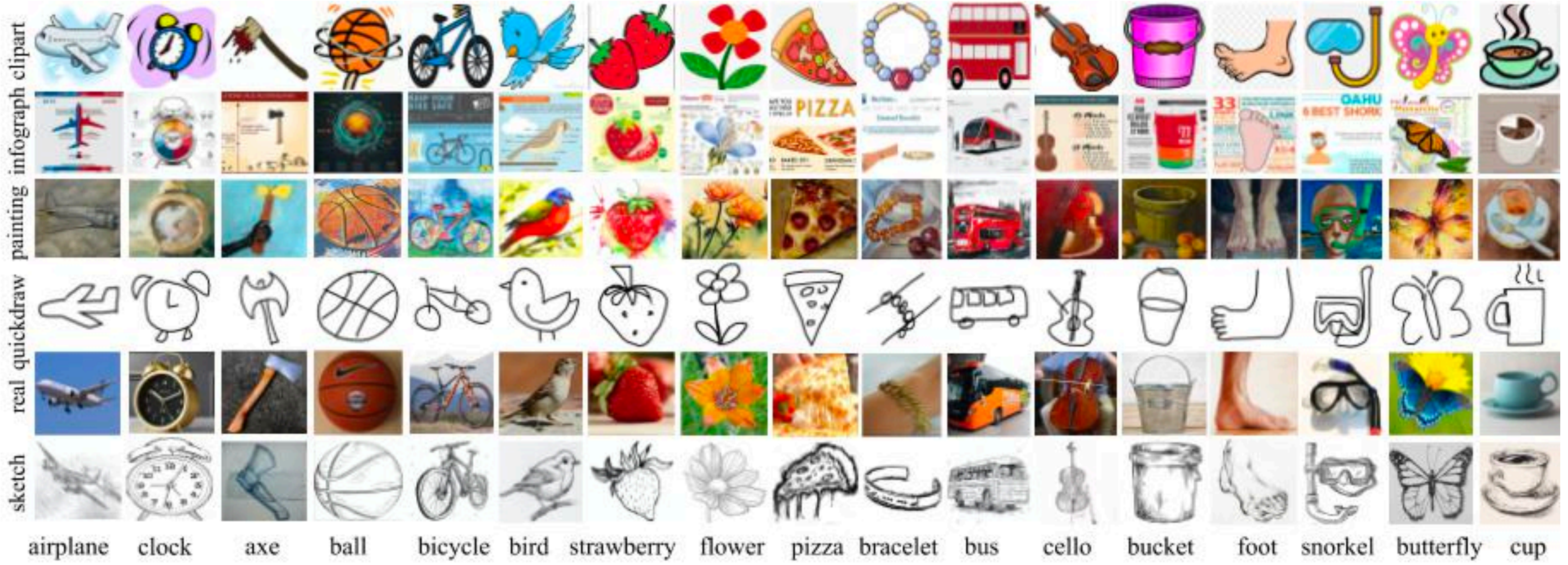
泛化能力

泛化能力（generalizability）



泛化能力

泛化能力（generalizability）



泛化误差上界

以二分类任务为例：

训练集 $\{(x_1, y_1), \ldots, (x_n, y_n)\}, \quad x \in R^D, \quad y \in \{+1, -1\}$

模型假设空间 $\mathbf{f} = \{f_1, \ldots, f_d\}$

训练误差 $\hat{R}(f) = \frac{1}{N} \sum_{n=1}^{N} I(y_n \neq f(x_n))$

测试误差的期望 $R(f) = E(x, y)[I(y \neq f(x_n)]$

泛化误差上界

Vapnik-Chervonenkis 维度

对任意一个函数，至少以概率$1-\delta$，以下不等式成立

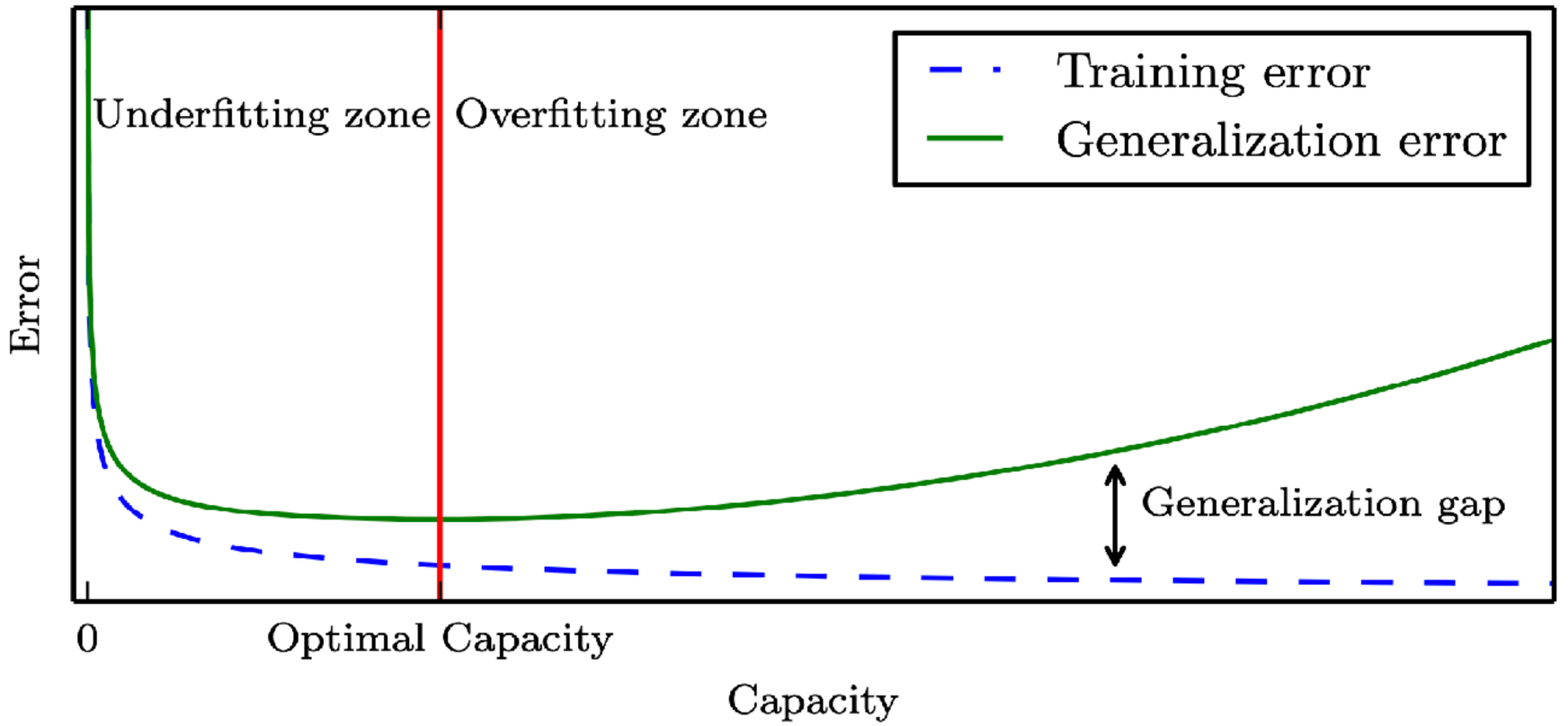
$$R(f) = \hat{R}(f) + \sqrt{\frac{1}{2N}\left(logd + log\frac{1}{\delta}\right)}$$

误差上界

- 训练误差小的模型，泛化误差上界也会越小
- 样本数量增大，泛化误差上界减小
- 假设空间越大，泛化误差上界减小

欠拟合、过拟合

从经验误差和泛化误差的角度：



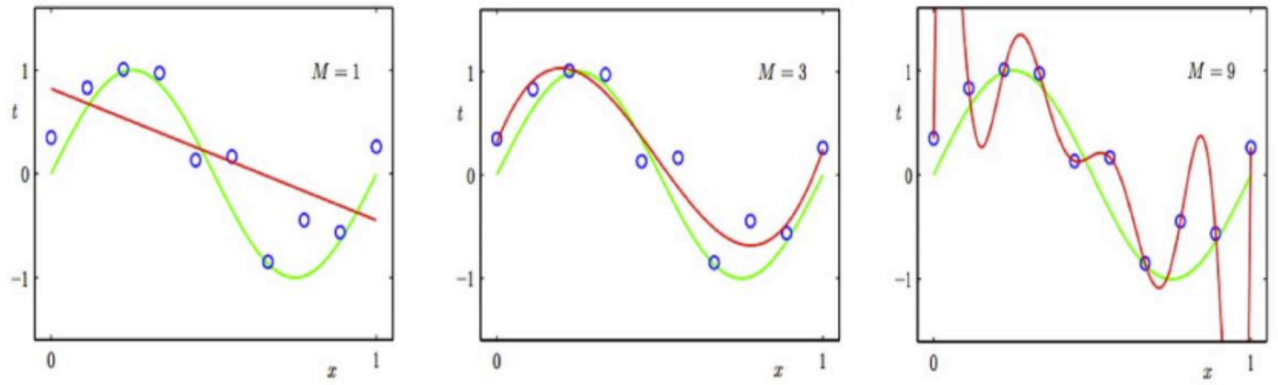
欠拟合：经验误差和泛化误差都非常高

过拟合：经验误差降低、泛化误差反升

欠拟合、过拟合

以回归和分类为例：

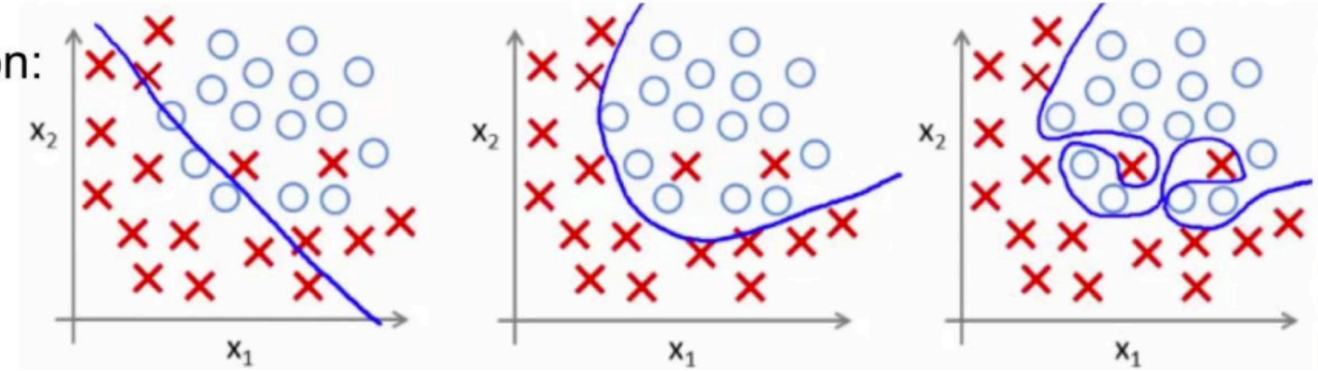
Regression:



predictor too inflexible:
cannot capture pattern

predictor too flexible:
fits noise in the data

Classification:



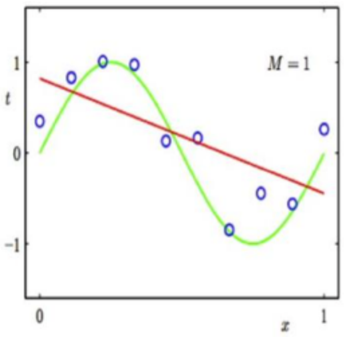
欠拟合、过拟合

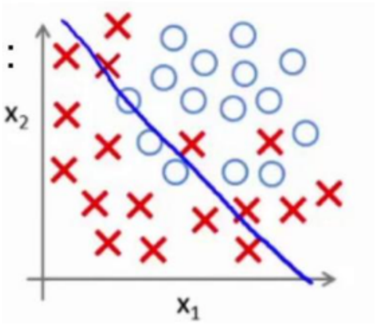
欠拟合：模型没有在训练集上取得足够低的误差

（模型表达能力太低或训练迭代次数不足）

改用表达能力更强的模型、

增加训练迭代次数

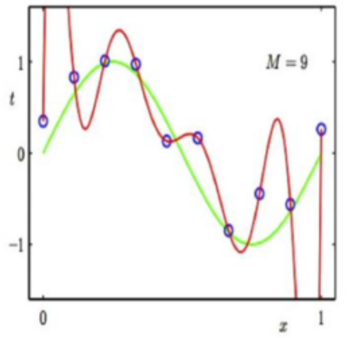


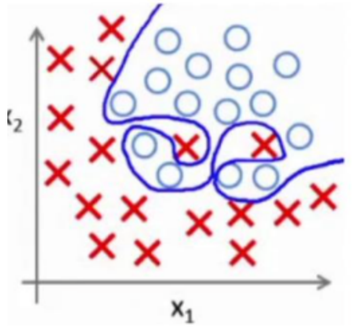


过拟合：训练误差和和测试误差之间的差距太大

（模型表达能力过强，训练样本不足导致拟合训练集噪声）

增加训练数据、参数正则化、早停止策略





正则化

正则化（Regularization）指对降低泛化误差而非训练误差的修改

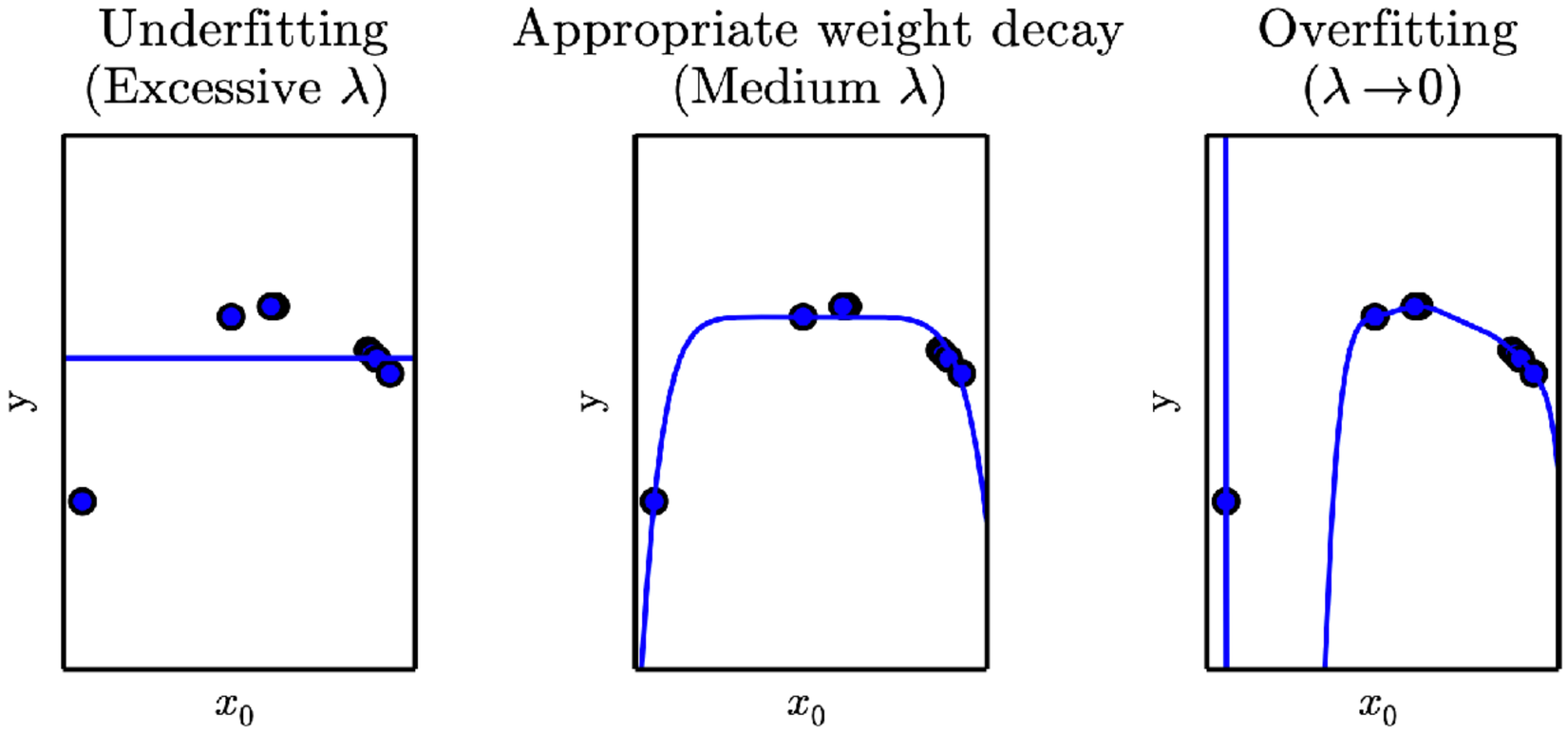
权重衰减（Weight Decay）

以训练九次多项式为例

$$y = \sum_{i}^{9} a_i x^i$$

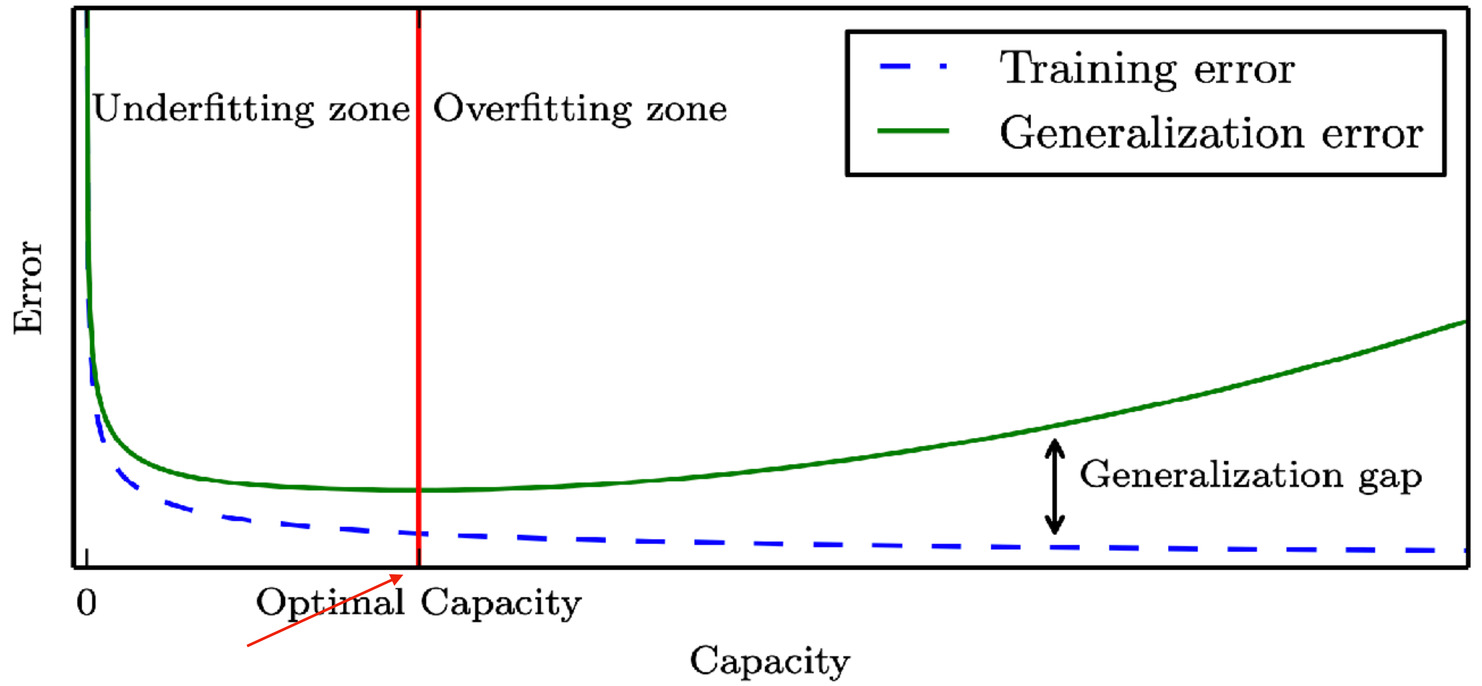
$$w = [a_0, a_1, \ldots, a_9]$$

$$J(w) = MSE_{train} + \lambda w^T w$$



早停止策略

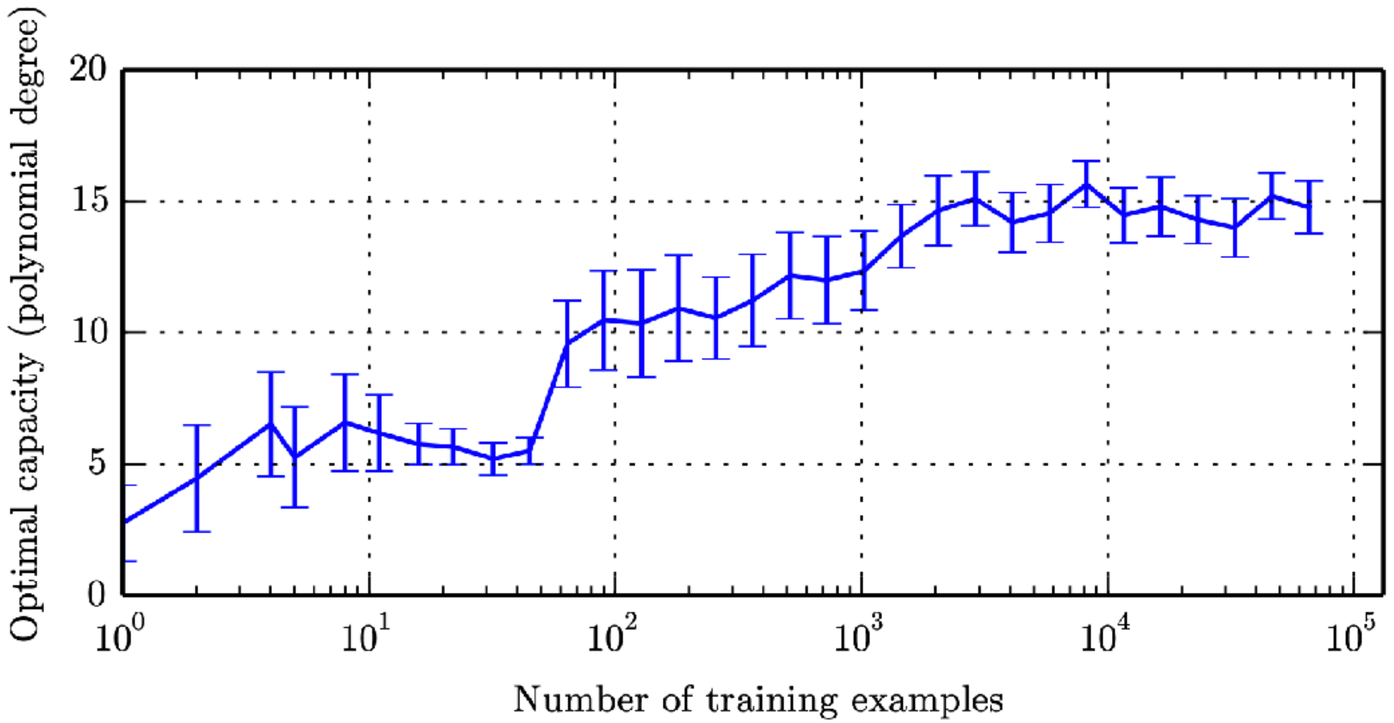
- 将数据划分为训练集和验证集，验证集用来估计误差
- 若训练集误差降低，验证集误差升高，则停止训练



停止

增大加训练数据

- 增加训练数据可以有效防止过拟合
- 大数据量可以抵消模型拟合能力过强的问题



偏差与方差(bias and variance)

偏差(bias)：期望输出$\bar{f}(x)$与真实标记y的偏离程度。

$$bias^2(x) = \left(\bar{f}(x) - y\right)^2$$

方差(variance)：x在训练集D上学得模型$f(x;D)$上的输出与期望输出$\bar{f}(x)$之间的变动导致的学习性能的变化。

$$var(x) = E_D[\left(f(x;D) - \bar{f}(x)\right)^2]$$

噪声：$\varepsilon^2 = E_D[(y_D - y)^2]$

偏差与方差(bias and variance)

假定噪声期望$E_D\left[y - y_D\right]$为0

$$
\begin{aligned}
E(f;D) &= E_D[(f(x;D) - y_D)^2] \\
&= E_D\left[(f(x;D) - \bar{f}(x) + \bar{f}(x) - y_D)^2\right] \\
&= E_D\left[\left(f(x;D) - \bar{f}(x)\right)^2\right] + E_D\left[\left(\bar{f}(x) - y_D\right)^2\right] + E_D\left[2\left(f(x;D) - \bar{f}(x)\right)\left(\bar{f}(x) - y_D\right)\right] \\
&= E_D\left[\left(f(x;D) - \bar{f}(x)\right)^2\right] + E_D\left[\left(\bar{f}(x) - y_D\right)^2\right] \\
&= var(x) + E_D\left[\left(\bar{f}(x) - y + y - y_D\right)^2\right] \\
&= var(x) + E_D\left[\left(\bar{f}(x) - y\right)^2\right] + E_D\left[\left(y - y_D\right)^2\right] + E_D\left[2(\bar{f}(x) - y)(y - y_d)\right] \\
&= var(x) + E_D\left[\left(\bar{f}(x) - y\right)^2\right] + E_D\left[\left(y - y_D\right)^2\right] \\
&= var(x) + bias^2(x) + \varepsilon^2
\end{aligned}
$$

偏差与方差(bias and variance)

$$E(f;D) = var(x) + bias^2(x) + \varepsilon^2$$

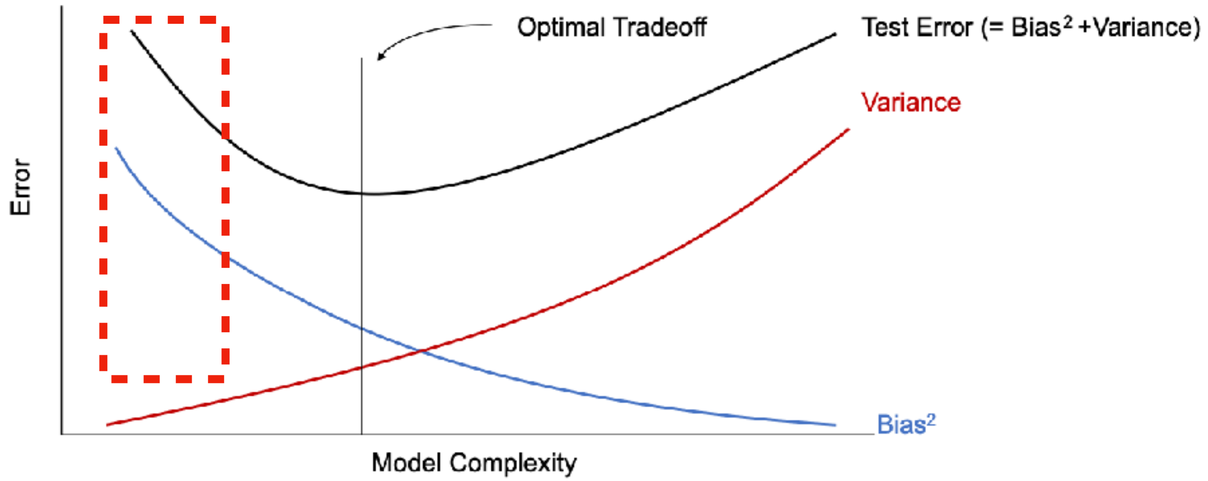
$var(x)$ 训练集的变动（数据扰动）导致的影响

$bias^2(x)$ 算法本身的拟合能力

ε^2 问题本身的难度

偏差与方差(bias and variance)

high bias
underfit



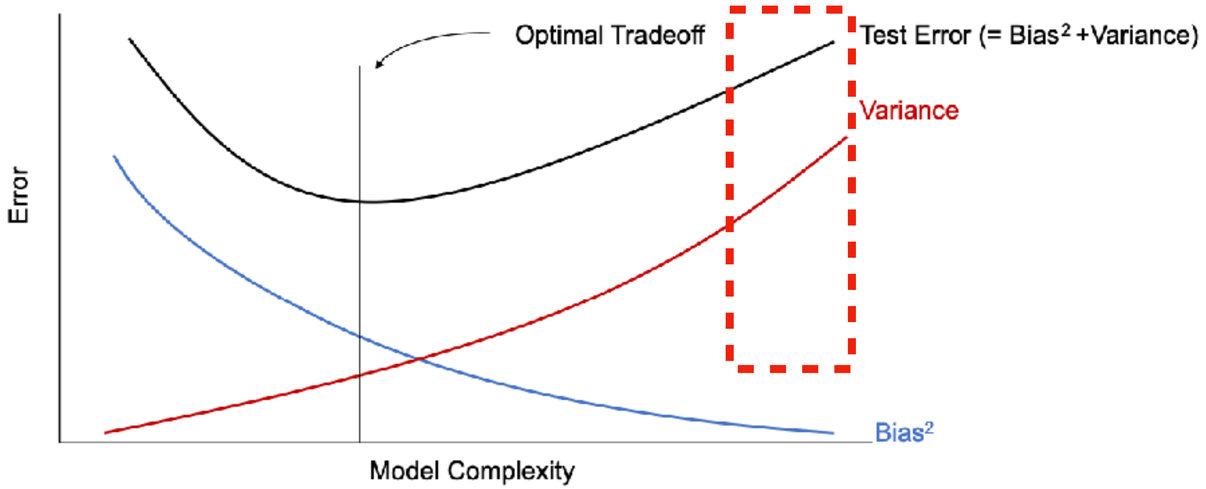
trade-off

complexity ⟷ fitting

偏差与方差(bias and variance)

high variance

overfit



trade-off

complexity ⟷ fitting

偏差与方差(bias and variance)