

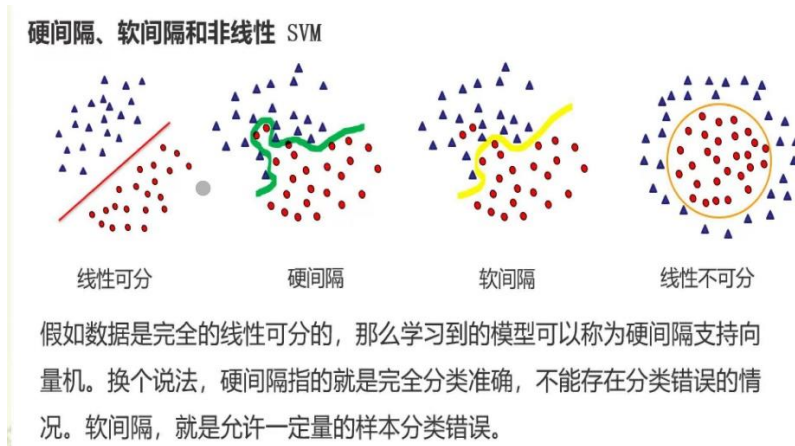
# 支持向量机

배경: 전통적인 통계식별 방법은 樣本이 무한대로 많을때만 성능이 보장되었는데, 统计学习理论 (STL) 연구는 유한한 樣本을 가지고 머신러닝 문제를 연구한다. 그리고 SVM은 통계학학습 이론에 기초한다.

전통적인 통계식별방법은 머신러닝을 진행할 때, 经验风险最小化를 강조하는데, 过学习問題を 발생시키고, 推广能力(예측 능력)를 저하시킨다. 하지만 SVM은 基于结构风险最小化准则的学习方法이고, 전통적인 학습방법보다 명확하게 우월하다.

支持向量机은 小样本, 非线性, 高维模式识别问题에서 우세하다. 그리고 线性可분할때는, 求解最后转化成二次规划问题的求解, 因此SVM的解是全局唯一的最优解。

数据가 线性可분인지 아닌지에 따라 模型의 성질이 바뀐다:



支持向量&间隔의 개념: 지지벡터는 그냥 超平面이랑 가장가까운 점을 기준으로 만든 平행하는 2개의 平面, 间隔는 支持向量두개간의 距离

假设超平面  $(w, b)$  能将训练样本正确分类, 即对于  $(x_i, y_i) \in D$ , 若  $y_i = +1$ , 则有  $w^T x_i + b > 0$ ; 若  $y_i = -1$ , 则有  $w^T x_i + b < 0$ . 令

$$\begin{cases} w^T x_i + b \geq +1, & y_i = +1; \\ w^T x_i + b \leq -1, & y_i = -1. \end{cases} \quad (6.3)$$

如图 6.2 所示, 距离超平面最近的这几个训练样本点使式(6.3)的等号成立, 它们被称为“支持向量”(support vector), 两个异类支持向量到超平面的距离之和为

$$\gamma = \frac{2}{\|w\|},$$

它被称为“间隔”(margin).

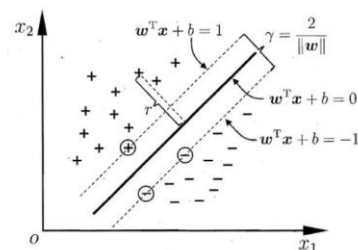


图 6.2 支持向量与间隔

자 이제 支持向量机가 뭔지 공부해보자: 支持向量机는 二分类模型, 定义在特征空间上的间隔最大的线性分类器。간격이 가장 크다는 점에서 感知机랑 차별점이 있다

线性可分支持向量机: 训练数据集가 线性可分해서, 간격을 최대화 해서 얻은 超平面에 상응하는 분류함수가 바로 线性可分支持向量机

종류:

- 1) 线性可分支持向量机: 数据线性可分, 硬间隔最大化, 学习一个线性分类器-硬间隔SVM
- 2) **线性支持向量机**(시험에 나왔음): 数据近似线性可分, 软间隔最大化, 学习一个线性分类器-软间隔SVM
- 3) 非线性支持向量机: 数据线性不可分, 用核技巧&软间隔最大化 - 非线性SVM。

이때 유용한 방법 이 등장: 拉格朗日乘子法.....!

用于求解带有约束条件的优化问题的数学方法。将约束条件引入到优化目标函数中, 从而将一个约束优化问题转化为一个无约束优化问题。

- ➔ 步骤: 构造拉格朗日函数, 结合目标函数和约束条件
- ➔ 对拉格朗日函数分别对变量和拉格朗日乘子法子求偏导并设为零
- ➔ 解方程组, 得到最优解

解的疏密性: 训练完成后, 大部分的训练样本都不需要保留, 最终模型仅与支持向量有关。중요하지 않은 포인트(오차가 크지 않은 데이터)들은 무시되므로 과적합을 방지할 수 있다.

线性不可分하면 어떻게 해요???

- ➔ 将样本从原始空间映射到一个**更高维**的特征空间, 使得样本在这个特征空间内线性可分。

一维 → 二维 / 二维 → 三维이런식으로 线性变换

# 核函数

核函数基本想法：不显示地设计核映射，而是设计核函数

Mercer定理：只要一个对称函数所对应的核矩阵半正定，则它就能作为核函数来使用。뭐라노

그니까 线性不可分할 때, 고차원 매핑을 하는데, 계산 비용이 엄청 올라갈 수 있으니, 명시적으로 수행하는게 아니라, 두 데이터 점간의 内积만으로 동일한 효과를 낸다, 그래서 연산 효율이 올라감.

常用核函数:

- 1) 线性核函数：不进行任何非线性变换，仅仅输入向量 $x$ 和 $x'$ 的内积(계산 빠르고 간단)

$$K(x, x') = \langle x, x' \rangle$$

- **定义**：线性核是最简单的核函数，它实际上不进行任何非线性变换，仅仅计算输入向量  $x$  和  $x'$  的内积。
- **适用场景**：当数据本身在原始空间中就可以线性分割时，使用线性核是最有效的选择。
- **优点**：计算速度快，适用于线性可分的数据集。
- **缺点**：无法处理复杂的非线性数据。

- 2) 多项式核函数：多项式核函数输入内积后进行多项式变换(非线性데이터에 좋음, 계산 복잡, 过拟合 위험)

$$K(x, x') = (\gamma \langle x, x' \rangle + r)^d$$

- **定义**：多项式核函数通过计算输入向量  $x$  和  $x'$  的内积并对其进行多项式变换。参数  $d$  是多项式的度数， $r$  是常数项， $\gamma$  控制内积的缩放。
- **适用场景**：适用于数据中存在非线性关系且数据维度较低的情况。
- **优点**：能够通过调整多项式的度数来拟合更复杂的数据结构。
- **缺点**：计算开销较大，容易过拟合，尤其是当度数  $d$  较大时。

- 3) PBF核函数(젤 자주 사용함)：计算欧氏距离进行高斯变换，使得数据点之间的相似度与距离的平方成负相关(복잡한非线性关系를 처리할 수 있음, 대부분 상황에 다 적용 가능, 参数调节 요구가 비교적 높은 편)

$$K(x, x') = \exp(-\gamma \|x - x'\|^2)$$

- **定义**：RBF 核函数（也叫高斯核）通过计算样本  $x$  和  $x'$  之间的欧氏距离并进行高斯变换，使得数据点之间的相似度与距离的平方成负相关。参数  $\gamma$  控制了数据点之间的影响范围。
- **适用场景**：当数据是非线性可分的，RBF 核非常适用，能够将数据映射到一个高维空间，使得数据在该空间中变得线性可分。
- **优点**：非常强大的非线性映射能力，能够处理复杂的决策边界。
- **缺点**：对参数  $\gamma$  和  $C$  非常敏感，需要调参；计算复杂度较高。

- 4) Sigmoid 核函数(거의 안씀): 输入向量的内积并应用双曲正切函数。(神经网络중의 激活函数 같은 특정한 곳에서만 사용)

$$K(x, x') = \tanh(\gamma \langle x, x' \rangle + r)$$

- **定义:** Sigmoid 核函数是一个类似于神经网络中激活函数的函数, 通过计算输入向量的内积并应用双曲正切函数 ( $\tanh$ )。参数  $\gamma$  控制了内积的影响,  $r$  是常数项。
- **适用场景:** Sigmoid 核在某些情况下可能会有效, 尤其是在某些结构化数据上, 但它通常不如 RBF 核常用。
- **优点:** 在某些特定问题上可能会表现良好, 尤其是当数据有类似神经网络中的激活函数的模式时。
- **缺点:** 比 RBF 核和多项式核更少使用, 通常较难调节。

核函数的选择 = SVM最大变数

文本数据 → 线性核函数

상황 애매 → 高斯核函数 (RBF核函数)

核函数的线性组合 仍为 核函数

核函数的直积仍为核函数

## 软间隔与正则化

0/1损失函数：最大化间隔的同时，让不满足约束的样本尽可能少。(쉽게 말해서 예측결과와 정확도를 측정하는 함수)

存在问题：0/1损失函数非凸，非连续，不易优化

그래서 引入“松弛变量”!!!!


每一个样本对应一个松弛变量，用来表示该样本不满足约束的程度

求解软间隔问题 → 构造拉格朗日函数，分别对变量求导，并令其为0.


正则化：머신러닝 모델이 과적합(Overfitting) 되는 것을 방지하기 위해 사용되는 기법

### □ 支持向量机器学习模型的更一般形式

$$\min_f \Omega(f) + C \sum_{i=1}^m l(f(x_i), y_i)$$

结构风险, 描述模型的某些性质

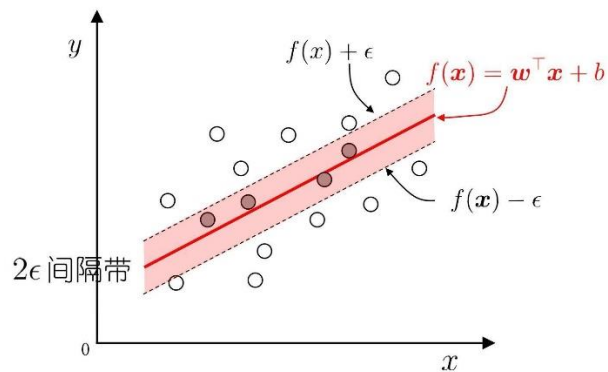
  

经验风险, 描述模型与训练数据的契合程度

여기서 사용된 정규화는 구조 위험 부분임, 모델의 복잡도를 제어하고 일반화를 돕기위한 추가 제약 조건을 제공하는 것.

## 支持向量回归

支持向量回归가 무엇이나, 允许模型输出和实际输出间存在 $2\epsilon$  的偏差。



전통적인 회귀모델은 기본적으로 모델의 输出 $f(x)$ 와 실제  $y$ 값의 차이로 손실을 계산했음. 두 개의 값이 완전히 같아야만 손실이 0이 되는거지. 이거와 다르게 支持向量回归 (SVR) 은  $\epsilon$ 의 편차가 있는걸 용인한다고 가정하고, 두개 값의 차의 절대값이  $\epsilon$ 일때만 손실을 계산함.

이렇게 하면 좋은점은  $2\epsilon$ 间隔带안에 있는 데이터들을 손실로 계산하지 않아서, 모델이 疏密性を 얻게 된다 → 防止过采样 (overfitting)

원래 문제를 形式化해서 svr도출해내면 이런 결과가 나온다.

$$f(\mathbf{x}) = \sum_{i=1}^m (\hat{\alpha}_i - \alpha_i) \mathbf{x}_i^T \mathbf{x} + b.$$

能使式(6.53)中的  $(\hat{\alpha}_i - \alpha_i) \neq 0$  的样本即为 SVR 的支持向量, 它们必落在  $\epsilon$ -间隔带之外. 显然, SVR 的支持向量仅是训练样本的一部分, 即其解仍具有稀疏性.

## Take Home Message

- 支持向量机的“最大间隔”思想
- 对偶问题及其解的稀疏性
- 通过向高维空间映射解决线性不可分的问题
- 引入“软间隔”缓解特征空间中线性不可分的问题
- 将支持向量的思想应用到回归问题上得到支持向量回归
- 将核方法推广到其他学习模型

## 支持向量机算法参数说明



### (2) 算法参数

使用 `scikit-learn` 机器学习库中的 `sklearn.svm` 类实现，其中 `SVC` 用来进行分类任务，`SVR` 用来进行数值回归任务。具体调用实例如下：

```
SVC(C=1.0, kernel='rbf', degree=3, gamma='auto', coef0=0.0, shrinking=True, probability=False, tol=0.001, cache_size=200, class_weight=None, verbose=False, max_iter=-1, decision_function_shape='ovr', random_state=None)
```

```
SVR(kernel='rbf', degree=3, gamma='auto', coef0=0.0, tol=0.001, C=1.0, epsilon=0.1, shrinking=True, cache_size=200, verbose=False, max_iter=-1)
```

参数	解释
<b>C</b>	惩罚系数，即对误差的宽容度。C 越高，说明越不能容忍出现误差，容易过拟合；C 越低，则容易欠拟合。
<b>kernel</b>	核函数，默认值为“rbf”，即高斯核函数。其他值为：线性函数“linear”，只能产生直线形状的分割超平面；多项式核函数“poly”，可构建出复杂形状的分割超平面。
<b>gamma</b>	是选择 rbf 函数作为 kernel 后该函数自带的一个参数，隐含的决定了数据映射到新的特征空间后的分布，gamma 越大，支持向量越少；gamma 值越小，支持向量越多。支持向量的个数会影响训练和预测的速度。



## 关于SVM使用的总结

57

下面是一些SVM普遍使用的准则:

$n$ 为特征数,  $m$ 为训练样本数。

(1)如果相较于 $m$ 而言,  $n$ 要大许多, 即训练集数据量不够支持我们训练一个复杂的非线性模型, 我们选用逻辑回归模型或者不带核函数的支持向量机。

(2)如果 $n$ 较小, 而且 $m$ 大小中等, 例如 $n$ 在 1-1000 之间, 而 $m$ 在10-10000之间, 使用高斯核函数的支持向量机。

(3)如果 $n$ 较小, 而 $m$ 较大, 例如 $n$ 在1-1000之间, 而 $m$ 大于50000, 则使用支持向量机会非常慢, 解决方案是创造、增加更多的特征, 然后使用逻辑回归或不带核函数的支持向量机。

## SVM方法的特点



- ① 非线性映射是SVM方法的理论基础, SVM利用内积核函数代替向高维空间的非线性映射;
- ② 对特征空间划分的最优超平面是SVM的目标, 最大化分类边际的思想是SVM方法的核心;
- ③ 支持向量是SVM的训练结果, 在SVM分类决策中起决定作用的是支持向量。
- SVM 是一种有坚实理论基础的新颖的小样本学习方法。它基本上不涉及概率测度及大数定律等, 因此不同于现有的统计方法。从本质上看, 它避开了从归纳到演绎的传统过程, 实现了高效的从训练样本到预报样本的“转导推理”(transductive inference), 大大简化了通常的分类和回归等问题。



# SVM方法的特点



- SVM 的最终决策函数只由少数的支持向量所确定, 计算的复杂性取决于支持向量的数目, 而不是样本空间的维数, 这在某种意义上避免了“维数灾难”。
- 少数支持向量决定了最终结果, 这不但可以帮助我们抓住关键样本、“剔除”大量冗余样本, 而且注定了该方法不但算法简单, 而且具有较好的“鲁棒”性。这种“鲁棒”性主要体现在:
  - ①增、删非支持向量样本对模型没有影响;
  - ②支持向量样本集具有一定的鲁棒性;
  - ③有些成功的应用中, SVM 方法对核的选取不敏感。

# SVM 应用



- 近年来SVM 方法已经在图像识别、信号处理和基因图谱识别等方面得到了成功的应用, 显示了它的优势。
- SVM 通过核函数实现到高维空间的非线性映射, 所以适合于解决本质上非线性的分类、回归和密度函数估计等问题。
- 支持向量方法也为样本分析、因子筛选、信息压缩、知识挖掘和数据修复等提供了新工具。