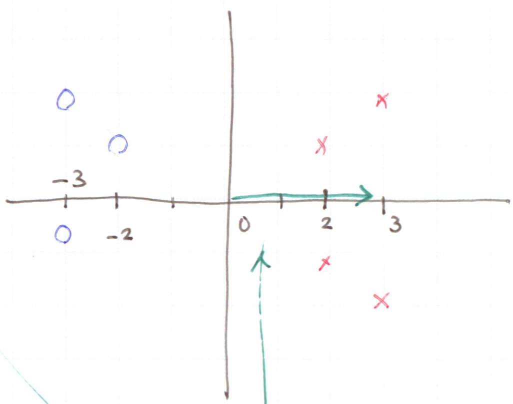
SVM : Mathematics Behind Large Margine Classification

$$\min_{\theta} \frac{1}{2}\sum_{j=1}^{n} \theta_j^2$$

s.t. $\|\theta\| \cdot p^{(i)} \geq 1$ if $y^{(i)} = 1$

such that $\|\theta\| \cdot p^{(i)} \leq -1$ if $y^{(i)} = 0$

$p^{(i)}$ は (符号つきの) $x^{(i)}$ から θ への 射影 (projection)



こういう向きのvectorで
長さ最小をもとめたい.
(条件をみたすこと)

$\theta) = (\theta_1, 0)$

$2 \cdot \theta_1 + y \cdot 0 \geq 1 \quad \therefore \theta_1 \geq \frac{1}{2}$

$-2 \cdot \theta_1 + y \cdot 0 \leq -1 \quad \therefore \theta_1 \geq \frac{1}{2}$

$\|\theta\|$ を最小化したいので

$\theta = (\frac{1}{2}, 0)$

$\|\theta\| = \frac{1}{2}$

SVM : Kernels

Non-linear Decision Boundary

Predict $y=1$ if $\theta_0+\theta_1x_1+\theta_2x_2+\theta_3x_1x_2+\theta_4x_1^2+\theta_5x_2^2+\cdots \geq 0$

kernel … Given x, landmark $l^{(1)}, l^{(2)}, l^{(3)}$ に近似した新しい feature を計算せよ.

$$\sum_{j=1}^{n}(x_j - l_j^{(1)})^2$$

$$f_1 = \underbrace{\text{similarity}}_{\text{kernel}}(x, l^{(1)}) = \underbrace{\exp\left(-\frac{\|x-l^{(1)}\|^2}{2\sigma^2}\right)}_{\text{Gaussian kernel}}$$

$$f_2 = \qquad {}^{(2)} \qquad\qquad {}^{(2)}$$

$$f_3 = \qquad {}^{(3)} \qquad\qquad {}^{(3)}$$

if $x \approx l^{(1)}$ $\quad f_1 \simeq \exp\left(-\frac{0^2}{2\sigma^2}\right) \simeq 1$

if x が $l^{(1)}$ から遠ければ

$$f_1 = \exp\left(-\frac{\text{大きな数}^2}{2\sigma^2}\right) \simeq 0$$

SVM: Kernels I の Quiz

x_1 は1次元の feature

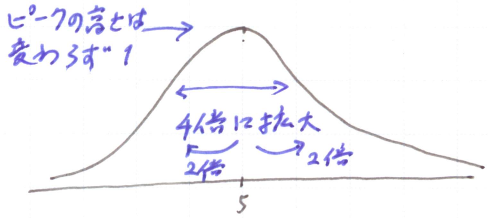
$l^{(1)} = 5$

$$f_1 = \exp\left(-\frac{\|x_1 - l^{(1)}\|}{2\sigma^2}\right)$$

$\sigma^2 = 1$ の時のグラフ

$y = e^{-\frac{x}{2}}$

$\sigma^2 = 4$ のときのグラフ



$y = e^{-\frac{x}{8}}$

SVM : Kernels II

Given x , $f_i = \text{similarity}(x, l^{(i)}) = \exp\left(-\frac{\|x - l^{(i)}\|^2}{2\sigma^2}\right)$

Predict $y=1$, if $\theta_0 + \theta_1 f_1 + \theta_2 f_2 + \theta_3 f_3 \geq 0$

$l^{(1)}, l^{(2)}, l^{(3)}, \ldots$ をどうやって見つければよいか？

SVM with Kernels

$(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})$ が与えられた時

$l^{(1)} = x^{(1)}, l^{(2)} = x^{(2)}, \ldots, l^{(m)} = x^{(m)}$ とする。

example x が与えられると

$f_1 = \text{similarity}(x, l^{(1)})$

$f_2 = \text{similarity}(x, l^{(2)})$

$\vdots$

まとめて $f = \begin{pmatrix} f_0 \\ f_1 \\ f_2 \\ \vdots \end{pmatrix}$, ただし $f_0 = 1$

training example $(x^{(i)}, y^{(i)})$ に対して

$f_1^{(i)} = \text{sim}(x^{(i)}, l^{(1)})$

$f_2^{(i)} = \text{sim}(x^{(i)}, l^{(2)})$

$\vdots$

$f_i^{(i)} = \text{sim}(x^{(i)}, l^{(i)}) = 1$ ∵同じものでー

$\vdots$

$f_m^{(i)} = \text{sim}(x^{(i)}, l^{(m)})$

まとめて $f^{(i)} = \begin{pmatrix} f_0^{(i)} \\ f_1^{(i)} \\ f_2^{(i)} \\ \vdots \\ f_m^{(i)} \end{pmatrix}$

1
これを除いて記述する場合がある.

与えられた x に対して、この f を計算する。 if $\theta^T f \geq 0$, predict $y=1$

Training (学習)

$n=m$ だそうだ

$$\min_{\theta} C \sum_{i=1}^{m} y^{(i)} cost_1 (\theta^T f^{(i)}) + (1-y^{(i)}) cost_0 (\theta^T f^{(i)}) + \frac{1}{2} \sum_{j=1}^{n} \theta_j^2$$

θ_0 は入っていない

$x^{(i)}$ の代わりに $f^{(i)}$ を使うのがポイント

SVM のパラメータ

$C \; (= \frac{1}{\lambda})$　　$C \to$ 大 ($\lambda \to$ 小)　　bias $\to$ 小
variance $\to$ 大　　(∵ θ_0 以外の $\theta_1 \sim \theta_n$ を小さくしようという影響が減るから)

$C \to$ 小 ($\lambda \to$ 大)　　bias $\to$ 大
variance $\to$ 小　　(∵ θ_0 以外の $\theta_1 \sim \theta_n$ を小さくしようとする影響が大きくなるから)

σ^2　　$\sigma^2 \to$ 大　　features f_i から離れていても正の値(?θ)をとる。
overfitting ぎみにはならない　ゆっくりと影響が減っていく。
bias $\to$ 大
variance $\to$ 小

$\sigma^2 \to$ 小　　上の逆で　bias $\to$ 小
variance $\to$ 大

[講義中の Quiz]

SVM で学習させたら overfit になった。どうすればよい

① C を増やす ($\lambda \to$ 小)　variance $\to$ 大　さらに overfit にする　×

② C を減らす ($\lambda \to$ 大)　θ_0 が相対的に大 ∴ bias 大　underfit へ　○

③ σ^2 を増やす。f_i から離れる影響がゆっくり $\to$ underfit へ　○

④ σ^2 を減らす。f_i から離れる影響が急激に。$\to$ overfit へ　×

SVMs in Practice

パラメータ C を選ぶ
kernel (similarity function) を選ぶ

(例) カーネルなし (linear kernel)
$y = 1$, if $\theta^T x \geq 0$ $\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n \geq 0$
nが大 mが小さいとき $x \in \mathbb{R}^{n+1}$

(例2) Gaussian kernel
$f_i = \exp\left(-\frac{\|x - l^{(i)}\|^2}{2\sigma^2}\right)$, where $l^{(i)} = x^{(i)}$ nが小 mが大 の時
σ^2 を選ぶ必要あり

Gaussian kernelを用いる前にfeature scalingを行うこと

全てのsimilarity functionがvalidなkernelとなるわけではない。
[SVMパッケージの最適化が正しく動作して、発散しない] → Mercer's Theorem を満たす必要がある

たくさんのoff-the-shelf kernelが利用可能である

polinominal kernel $(X^T l + \text{定数})^{\text{次数}}$

[講義中のQuiz]
C, σ^2のようなkernelのパラメータを選ぶ時、どうやって選ぶか?

① training dataでよい結果を出すもの
② cross validation data ← ○ パラメータはcross validation dataで決定する。
③ test data
④ SVM marginが最大のもの

SVMs in Practice
Multi-class classification

Multi-class classification 機能を持っているSVMパッケージも多い。

一方、one-vs-all 手法を用いることもできる。

Logistic regression vs. SVMs
featureの数 n とデータセットの数 m を比較して

n → 大、logistic regression
kernelなしの (linear kernelの) SVM } を使うべき
n = 10,000
m = 10 ~ 1000

n → 小、m → 中、Gaussian kernel の SVMを使うべき
n = 1 ~ 1000、m = 10 ~ 10,000

n → 小、m → 大　もっと多くのfeatureを追加して
n = 1 ~ 1000　m = 50,000 以上
logistic regression
kernelなしの (linear kernelの) SVM } を使うべき

上記の全ての場合で neural network はうまく動作するが、学習に時間がかかるかもしれない。

日本総研
The Japan Research Institute, Limited