# 半教師付き学習（16章）と転移学習（18章）

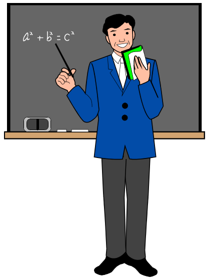
杉山将・本多淳也
sugi@k.u-tokyo.ac.jp, jhonda@k.u-tokyo.ac.jp
http://www.ms.k.u-tokyo.ac.jp

# 教師付き分類

- 教師付き分類:クラスラベル付きの訓練データから学習

$$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$$

- 精度の良い学習結果を得るためには,多くの訓練データが必要
- しかし,ラベル付きデータの収集にコスト(人手など)がかかる場合,多数のラベル付きデータを集められない




# 講義の流れ



1. 半教師付き学習（16章）
2. 転移学習（18章）

# 半教師付き学習

- ラベルなしのデータは簡単に手に入ることがある

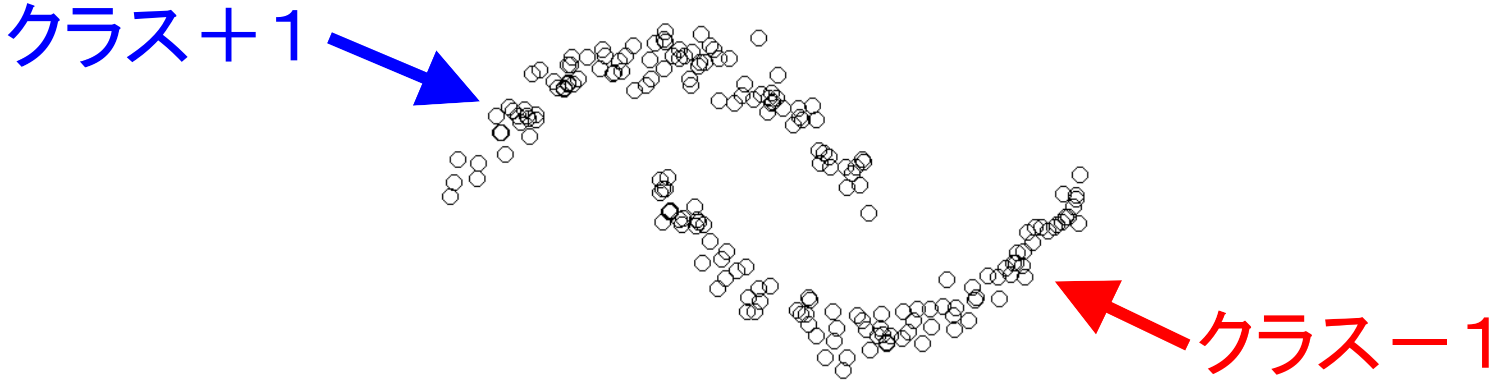
$$\{\boldsymbol{x}_i\}_{i=n+1}^{n+n'}$$

  - 顔認識：顔画像は容易に大量入手できるが，ラベル（性別，年齢など）付けは人手が必要
  - ウェブページ分類：ウェブページは自動的に大量入手できるが，ラベル（政治，芸能など）付けは人手が必要
- 半教師付き学習：ラベルなしデータも活用し，学習の精度を向上させる

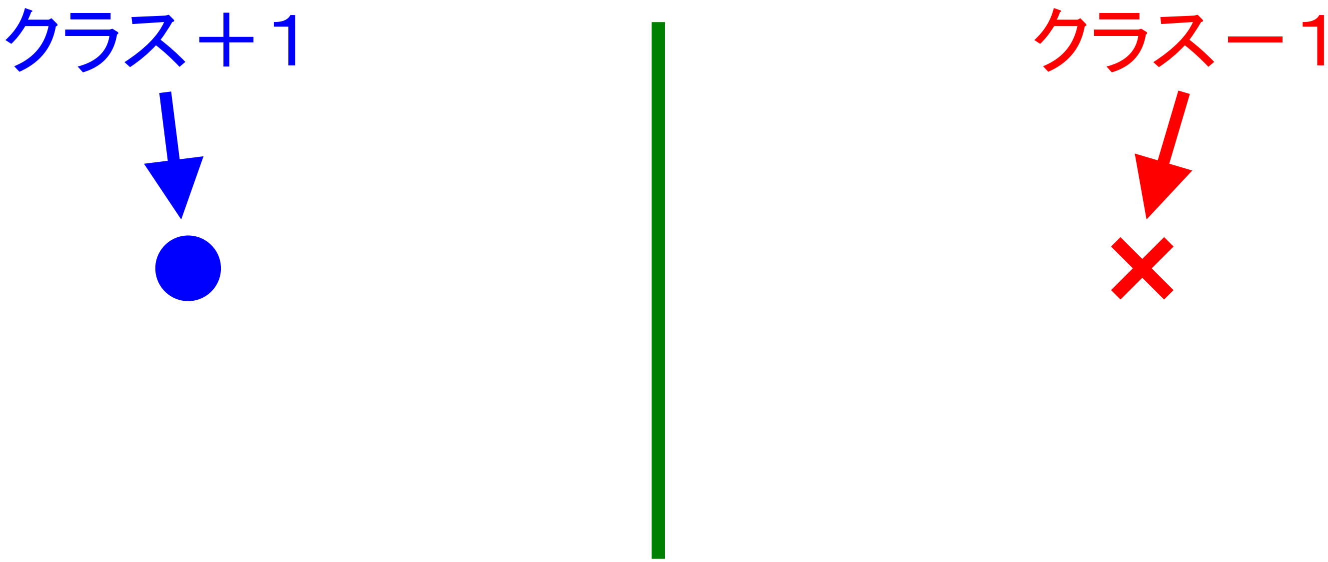
# 半教師付き学習の大前提



- 何も仮定をおかなければ，ラベルなしデータは一般に役に立たない
- 典型的な仮定：
  - 同じ“かたまり”に属するデータは，同じラベルを持つ

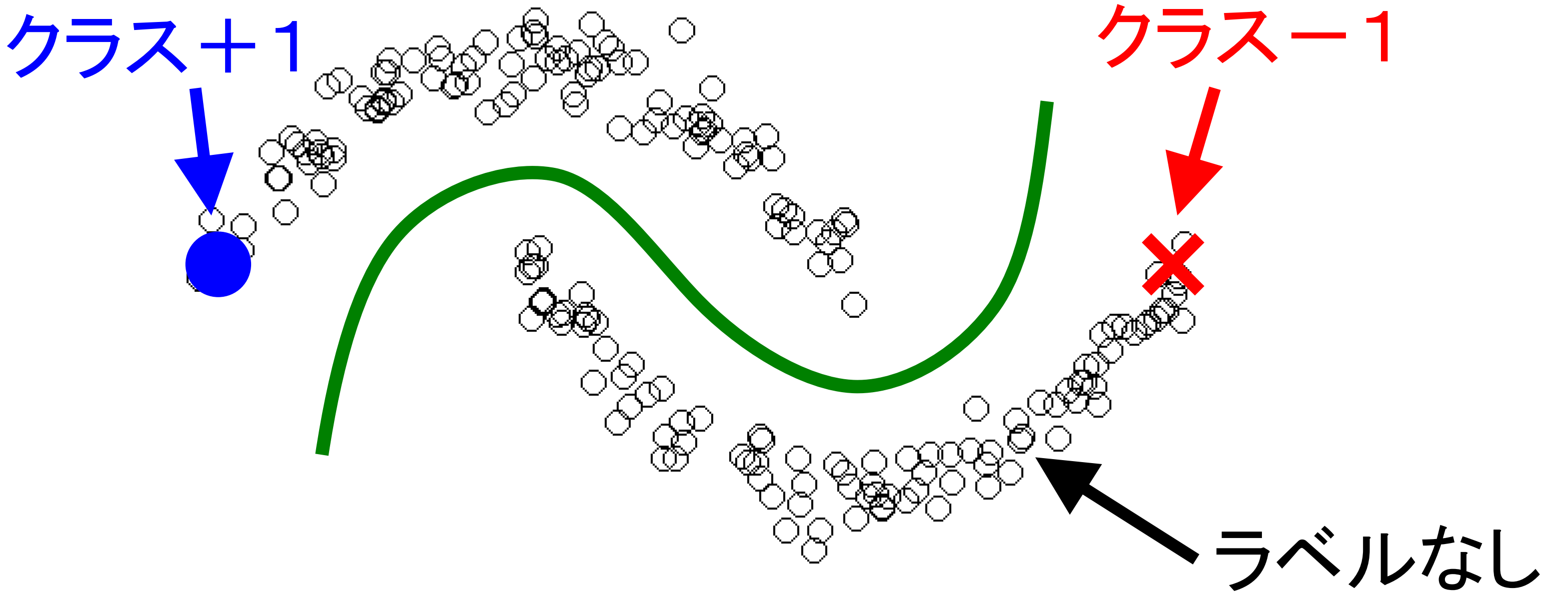


# 教師付き分類の場合




真ん中で分けるのが自然

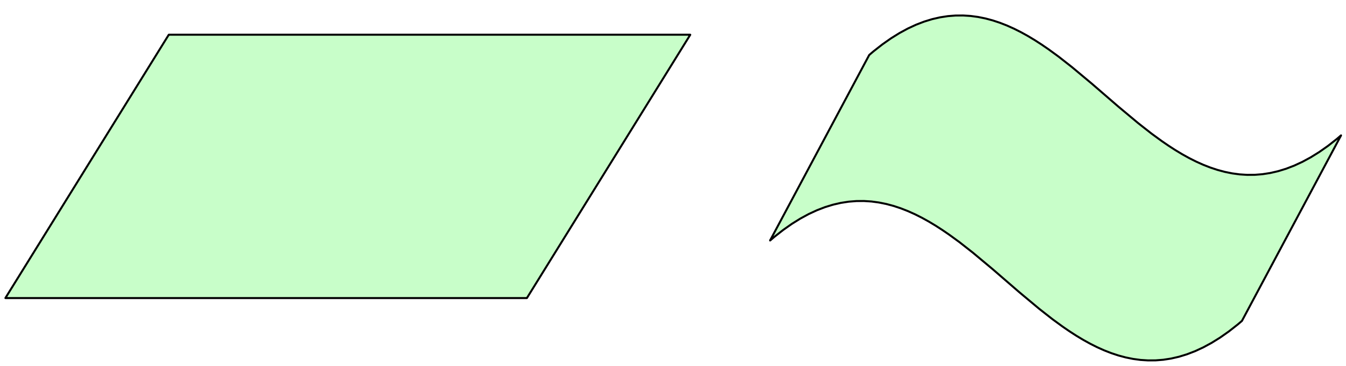
# 半教師付き学習の場合




ラベルなしデータがなす領域に
沿って分けるのが自然（多様体仮定）

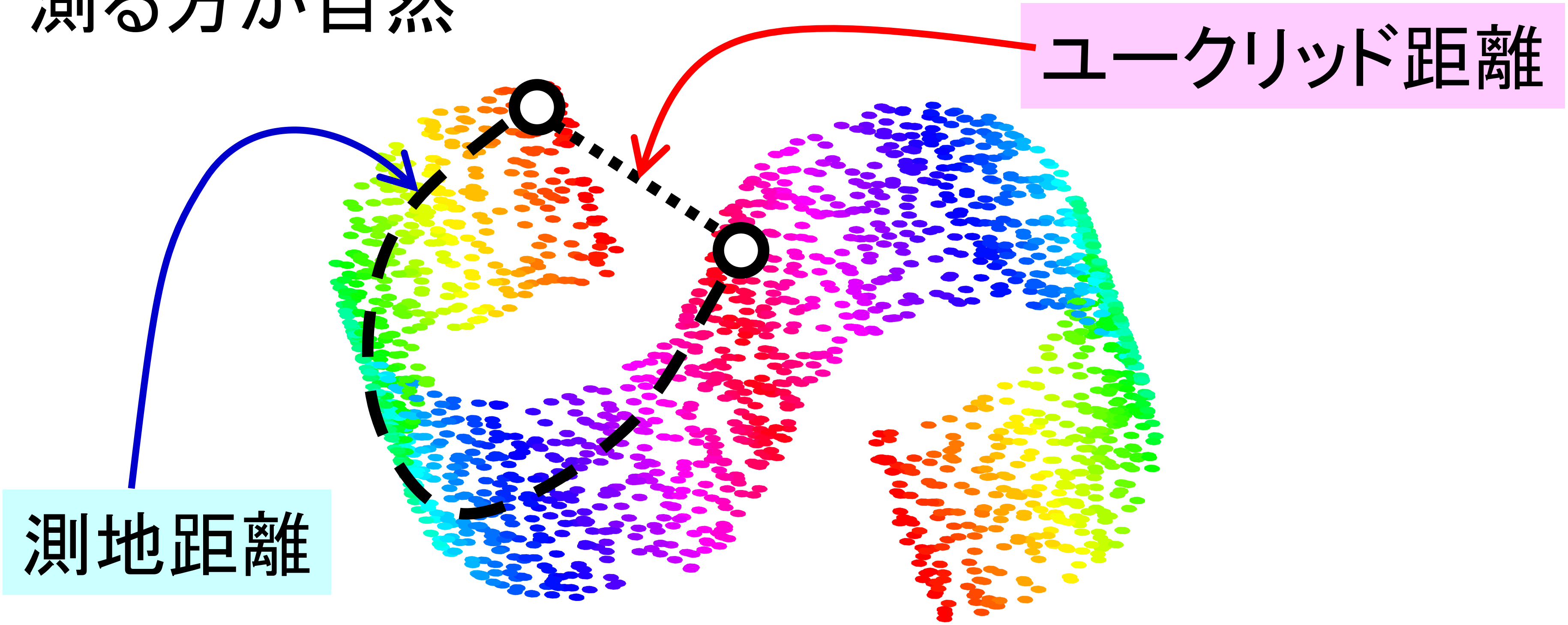
# 多様体とは

- 数学的な定義:
  - 局所的にユークリッド空間とみなせる空間
- 機械学習的な解釈:
  - 線形空間の非線形拡張



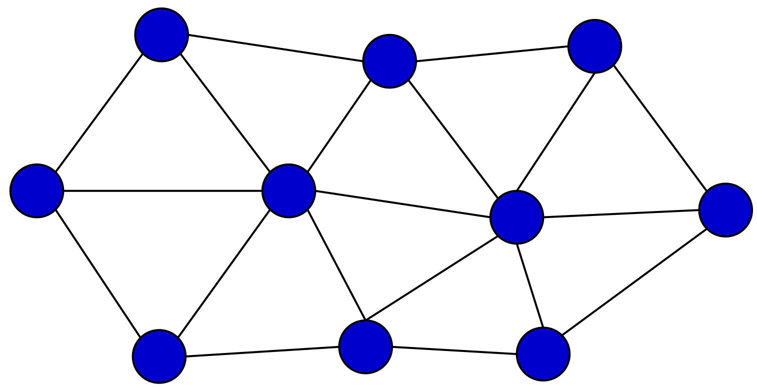
# 多様体を考える利点

- 高次元データは，空間全体に分布するのでなく，低次元の多様体上に分布することが多い．
- ユークリッド距離よりも，多様体上の測地線に沿った距離（多様体上での最短距離）で近さを測る方が自然



# 近傍グラフ

- 多様体は連続的な集合
- 与えられるデータは離散的な集合
- 多様体を，データから作られる近傍グラフで近似



$$W_{i,i'} = \begin{cases} 1 : \boldsymbol{x}_i \text{と} \boldsymbol{x}_{i'} \text{が } k \text{ 近傍} \\ 0 : \text{それ以外} \end{cases}$$

ガウスカーネルを使うこともある

# 正則化に基づく半教師付き学習

- ラベル付き訓練データ: $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$
- ラベルなし訓練データ: $\{\boldsymbol{x}_i\}_{i=n+1}^{n+n'}$

$$\underbrace{\sum_{i=1}^{n}\Big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i\Big)^2}_{\text{訓練出力に対する適合の良さ}} + \underbrace{\lambda\|\boldsymbol{\theta}\|^2}_{\text{正則化}} + \underbrace{\nu\sum_{i,i'=1}^{n+n'} W_{i,i'}\Big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - f_{\boldsymbol{\theta}}(\boldsymbol{x}_{i'})\Big)^2}_{\text{ラプラス正則化（後述）: 近傍の入力点間の出力の滑らかさ}}$$

$$W_{i,i'} = \begin{cases} 1 & : \boldsymbol{x}_i \text{と} \boldsymbol{x}_{i'} \text{が } k \text{ 近傍} \\ 0 & : \text{それ以外} \end{cases}$$

# 数学演習

- ラプラス行列 $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$ および $a_1, a_2, \ldots, a_m \in \mathbb{R}$ について次式を証明せよ

（$\boldsymbol{W}$ は対称行列）

$$\sum_{i,i'=1}^{m} W_{i,i'}(a_i - a_{i'})^2 = 2\sum_{i,i'=1}^{m} L_{i,i'}a_i a_{i'}$$

$$\boldsymbol{D} = \mathrm{diag}\left(\sum_{i=1}^{m} W_{1,i}, \ldots, \sum_{i=1}^{m} W_{m,i}\right)$$

# 解答例



$$\sum_{i,i'=1}^{m} W_{i,i'}(a_i - a_{i'})^2$$

$$= 2\sum_{i,i'=1}^{m} W_{i,i'} a_i^2 - 2\sum_{i,i'=1}^{m} W_{i,i'} a_i a_{i'}$$

$$= 2\sum_{i=1}^{m} D_{i,i} a_i^2 - 2\sum_{i,i'=1}^{m} W_{i,i'} a_i a_{i'}$$

$$= 2\sum_{i,i'=1}^{m} D_{i,i'} a_i a_{i'} - 2\sum_{i,i'=1}^{m} W_{i,i'} a_i a_{i'}$$

$$= 2\sum_{i,i'=1}^{m} L_{i,i'} a_i a_{i'}$$

# ラプラス正則化の変形

$$\sum_{i,i'=1}^{m} W_{i,i'}(a_i - a_{i'})^2 = 2\sum_{i,i'=1}^{m} L_{i,i'}a_i a_{i'}$$

- 線形モデル$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^\top \phi(\boldsymbol{x}) = \phi(\boldsymbol{x})^\top \boldsymbol{\theta}$を用いると

$$\sum_{i,i'=1}^{n+n'} W_{i,i'}\Big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - f_{\boldsymbol{\theta}}(\boldsymbol{x}_{i'})\Big)^2$$

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_1(\boldsymbol{x}_1) & \cdots & \phi_b(\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_1(\boldsymbol{x}_{n+n'}) & \cdots & \phi_b(\boldsymbol{x}_{n+n'}) \end{pmatrix}$$

$$= \sum_{i,i'=1}^{n+n'} \boldsymbol{\theta}^\top \phi(\boldsymbol{x}_i) L_{i,i'} \phi(\boldsymbol{x}_{i'})^\top \boldsymbol{\theta}$$

$$= \boldsymbol{\theta}^\top \left( \sum_{i,i'=1}^{n+n'} \phi(\boldsymbol{x}_i) L_{i,i'} \phi(\boldsymbol{x}_{i'})^\top \right) \boldsymbol{\theta} = \boldsymbol{\theta}^\top \boldsymbol{\Phi}^\top \boldsymbol{L} \boldsymbol{\Phi} \boldsymbol{\theta}$$

# ラプラス正則化最小二乗分類の解の求め方

$$\widehat{\boldsymbol{\theta}} = \mathop{\mathrm{argmin}}_{\boldsymbol{\theta}} \left[ \sum_{i=1}^{n} \Big( f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i \Big)^2 + \lambda \|\boldsymbol{\theta}\|^2 + \nu \sum_{i,i'=1}^{n+n'} W_{i,i'} \Big( f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - f_{\boldsymbol{\theta}}(\boldsymbol{x}_{i'}) \Big)^2 \right]$$

- 線形モデル $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^\top \phi(\boldsymbol{x})$ に対して

$$\widehat{\boldsymbol{\theta}} = \mathop{\mathrm{argmin}}_{\boldsymbol{\theta}} \left[ \|\widetilde{\boldsymbol{\Phi}}\boldsymbol{\theta} - \boldsymbol{y}\|^2 + \lambda \|\boldsymbol{\theta}\|^2 + 2\nu \boldsymbol{\theta}^\top \boldsymbol{\Phi}^\top \boldsymbol{L} \boldsymbol{\Phi} \boldsymbol{\theta} \right]$$

$$= (\widetilde{\boldsymbol{\Phi}}^\top \widetilde{\boldsymbol{\Phi}} + \lambda \boldsymbol{I} + 2\nu \boldsymbol{\Phi}^\top \boldsymbol{L} \boldsymbol{\Phi})^{-1} \widetilde{\boldsymbol{\Phi}}^\top \boldsymbol{y}$$

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_1(\boldsymbol{x}_1) & \cdots & \phi_b(\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_1(\boldsymbol{x}_{n+n'}) & \cdots & \phi_b(\boldsymbol{x}_{n+n'}) \end{pmatrix} \qquad \widetilde{\boldsymbol{\Phi}} = \begin{pmatrix} \phi_1(\boldsymbol{x}_1) & \cdots & \phi_b(\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_1(\boldsymbol{x}_n) & \cdots & \phi_b(\boldsymbol{x}_n) \end{pmatrix}$$

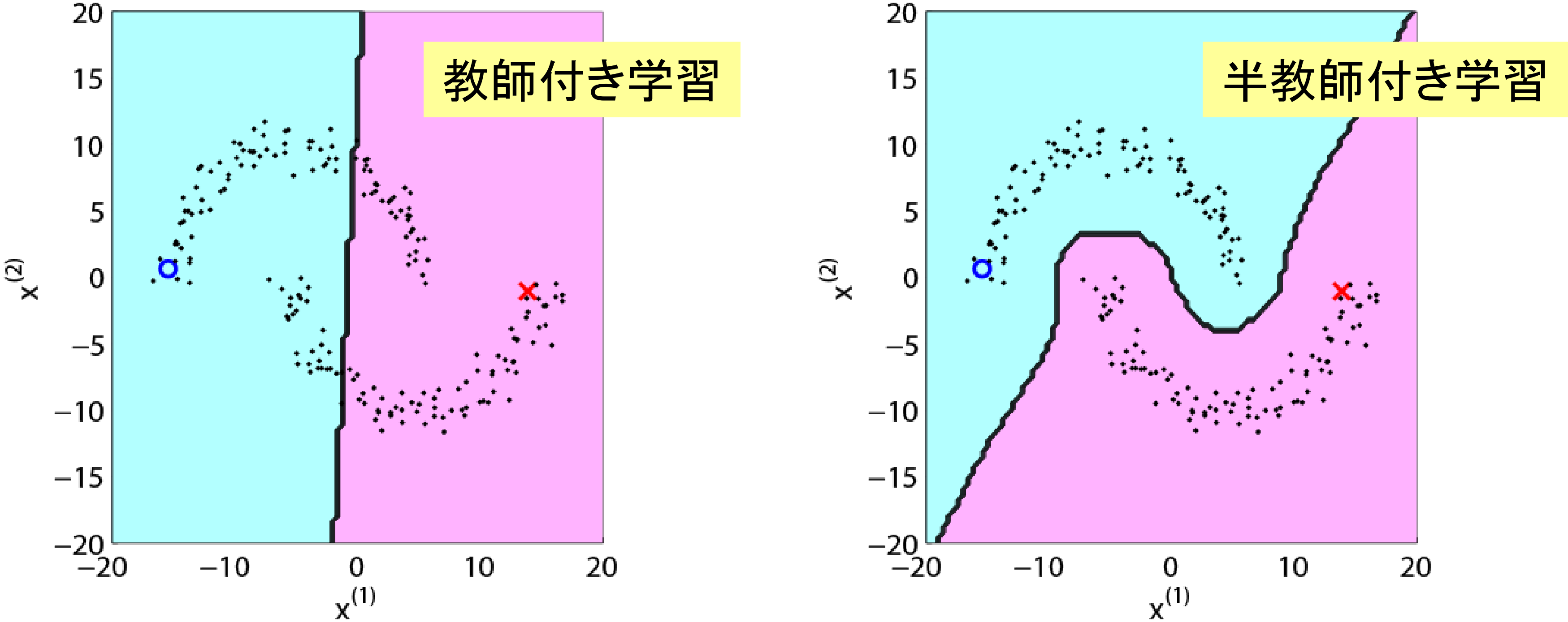
$$\boldsymbol{y} = (y_1, \ldots, y_n)^\top$$

# 実行例

- ガウスカーネルモデル

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{n+n'} \theta_j \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_j\|^2}{2h^2}\right)$$

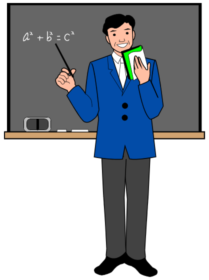
に対するラプラス正則化最小二乗分類




- ラプラス正則化により，ラベルなしデータの多様体に沿った分離境界が得られている

# 半教師付き学習：まとめ

- ラベル付きデータは大量に集めることが困難
- ラベルなしデータは容易に得られる
- 同じ「かたまり」に属するデータは同じラベルを持つという仮定のもと，ラベルなしデータを活用
  - グラフ・ラプラス行列を用いて，ラベルなしデータがなす多様体にそってラベルを伝播
- ラプラス正則化は，最小二乗分類以外の分類法や，回帰にも適用可能



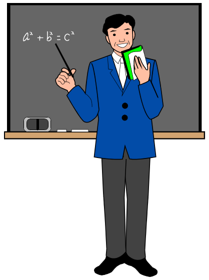
# 講義の流れ



1. 半教師付き学習（16章）
2. 転移学習（18章）
   A) 共変量シフト
   B) クラスバランス変化

# 転移学習 (Transfer Learning)

- 教師付き学習の大前提: 訓練標本とテスト標本が同じ確率分布に従う
  - この仮定のもと，訓練標本から関数を学習し，テスト入力に対する出力を予測
- 応用問題によっては，訓練標本とテスト標本の確率分布が異なる事がある
- そのような状況でも，訓練標本から得られる知見を転移させて，テスト入力に対する出力を予測したい



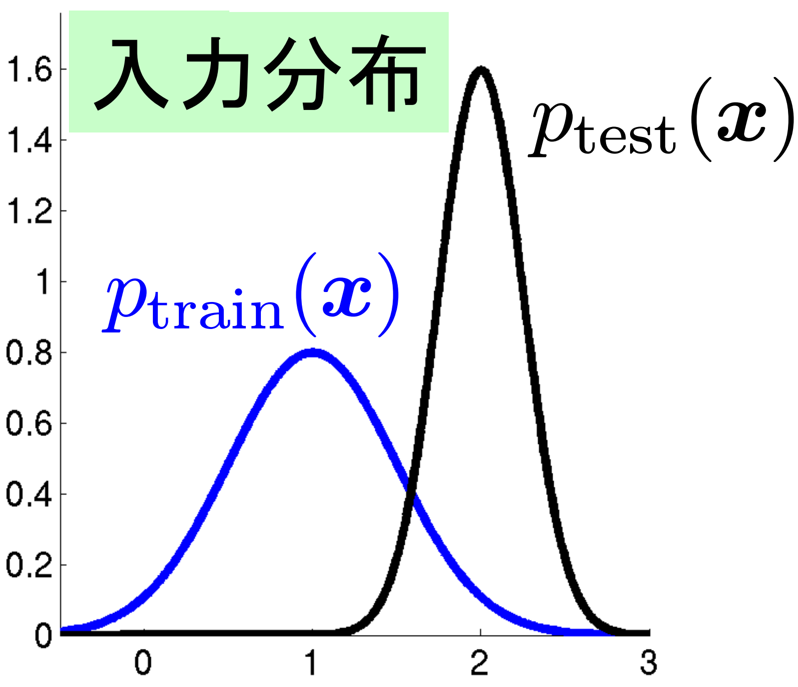
# 講義の流れ

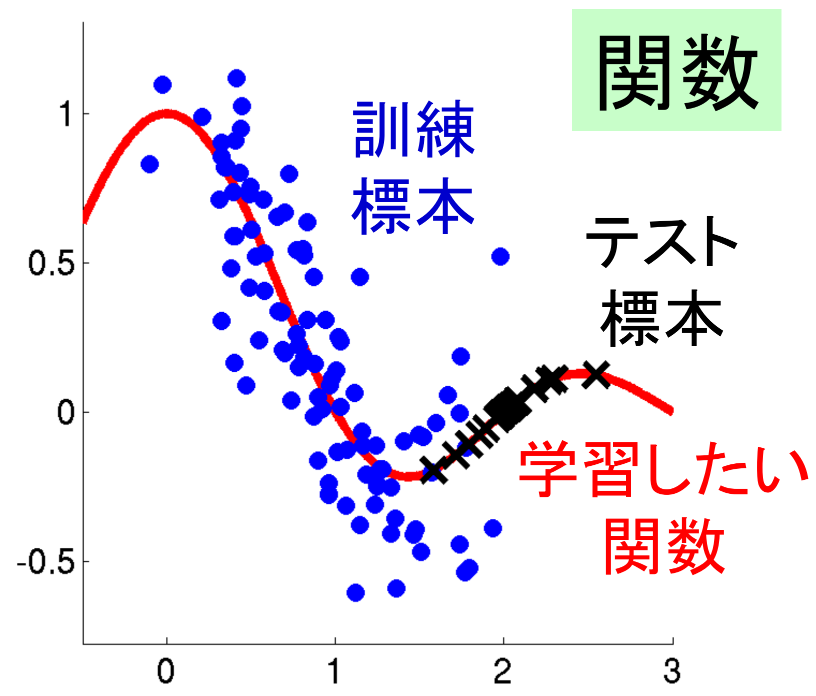


1. 半教師付き学習（16章）
2. 転移学習（18章）
   A) 共変量シフト
   B) クラスバランス変化

# 共変量シフト適応

- 共変量:入力の別名
- 共変量シフト:訓練時とテスト時で入力分布が変化するが，入出力関数は変わらない
- 外挿問題が典型的な例



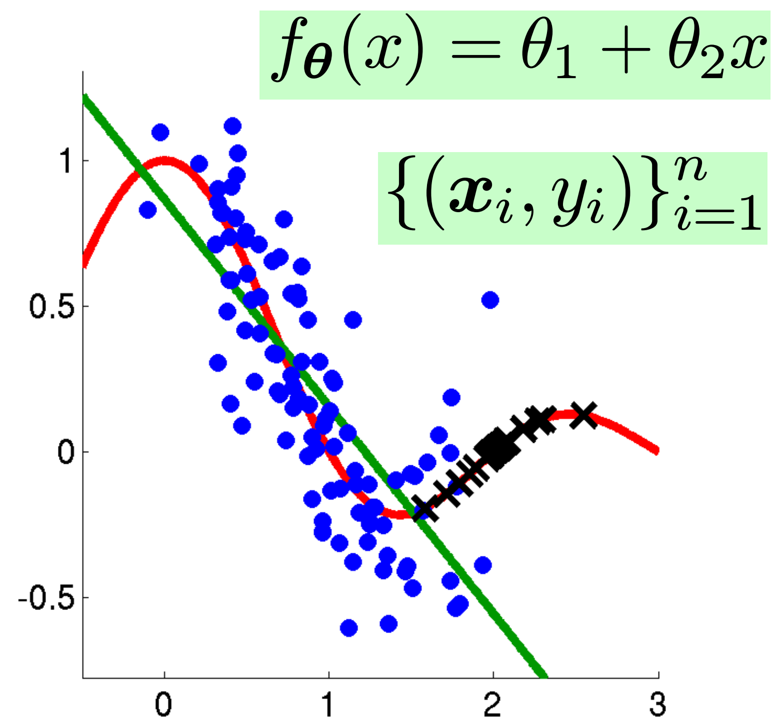
# 共変量シフトの例

- 医療データ:
  - 訓練データ: 何らかの理由で検査を受けることになった患者の診断結果
  - テストデータ: 人間ドック等を受けた一般人の検査データ
- スパムフィルタ:
  - 訓練データ: データベースに蓄積されたスパム/正常メール
  - テストデータ: 個々のユーザーに応じたスパム/正常メール

# 最小二乗法

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \Big( f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i \Big)^2$$

- 通常の設定では，最小二乗法は一致性を持つ．つまり，$n \to \infty$ で(モデル内で)最適な解が求まる
- しかし共変量シフト下では，モデルが真の関数を含まない限り最小二乗法は一致性をもたない



# 最小二乗法の一致性：大数の法則

- 標本平均は真の期待値に収束：

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{loss}_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \longrightarrow \int \mathrm{loss}_{\boldsymbol{\theta}}(\boldsymbol{x})p_{\mathrm{train}}(\boldsymbol{x})d\boldsymbol{x}$$

$$\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} p_{\mathrm{train}}(\boldsymbol{x})$$

- 共変量シフトがなければ，このことから一致性が保証される
- 訓練データ $\{\boldsymbol{x}_i\}_{i=1}^{n}$ を用いて，テストデータに対する期待値を最小にしたい！

$$\int \mathrm{loss}_{\boldsymbol{\theta}}(\boldsymbol{x})p_{\mathrm{test}}(\boldsymbol{x})d\boldsymbol{x}$$

# 重点(importance)サンプリング

- 重要度:テスト入力と訓練入力の密度の比

$$\frac{p_{\mathrm{test}}(\boldsymbol{x})}{p_{\mathrm{train}}(\boldsymbol{x})}$$

- 通常の平均損失のかわりに重要度重み付き平均を考えると，大数の法則により

$$\frac{1}{n}\sum_{i=1}^{n}\frac{p_{\mathrm{test}}(\boldsymbol{x}_i)}{p_{\mathrm{train}}(\boldsymbol{x}_i)}\mathrm{loss}_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \qquad \boldsymbol{x}_i \overset{\mathrm{i.i.d.}}{\sim} p_{\mathrm{train}}(\boldsymbol{x})$$

$$\longrightarrow \int \frac{p_{\mathrm{test}}(\boldsymbol{x})}{p_{\mathrm{train}}(\boldsymbol{x})}\mathrm{loss}_{\boldsymbol{\theta}}(\boldsymbol{x})p_{\mathrm{train}}(\boldsymbol{x})d\boldsymbol{x}$$

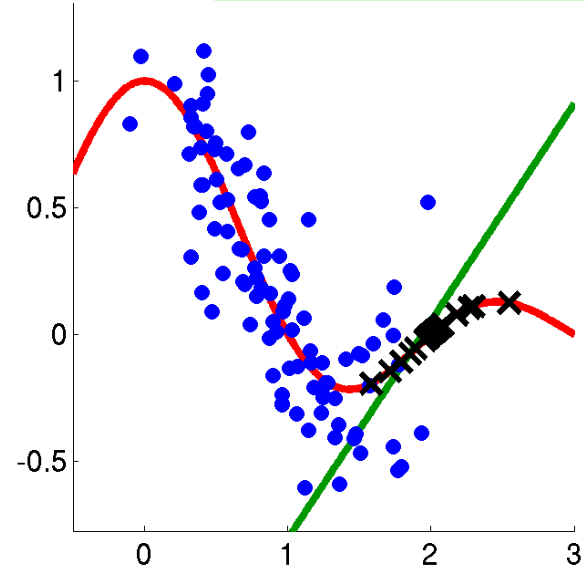
$$= \int \mathrm{loss}_{\boldsymbol{\theta}}(\boldsymbol{x})p_{\mathrm{test}}(\boldsymbol{x})d\boldsymbol{x}$$

# 重要度重み付き最小二乗法

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \frac{p_{\mathrm{test}}(\boldsymbol{x}_i)}{p_{\mathrm{train}}(\boldsymbol{x}_i)} \Big( f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i \Big)^2$$

$$f_{\boldsymbol{\theta}}(x) = \theta_1 + \theta_2 x$$

$$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$$




- 重要度重み付き最小二乗法は
共変量シフト下でも一致性を持つ
- 重要度重み付けは多くの学習法に適用できる:
  - サポートベクトルマシン
  - ロジスティック回帰
  - 条件付き確率場など

# 重要度の推定



- 重要度重み付き最小二乗法では，重要度

$$\frac{p_{\mathrm{test}}(\boldsymbol{x}_i)}{p_{\mathrm{train}}(\boldsymbol{x}_i)}$$

  が必要
- 単純な方法：確率密度 $p_{\mathrm{train}}(\boldsymbol{x})$と $p_{\mathrm{test}}(\boldsymbol{x})$を訓練データとテストデータからそれぞれ推定
- しかし，推定した確率密度での割り算は推定誤差を増幅させてしまう
- 重要度を直接推定したい

# 最小二乗重要度推定法

- データ:
  - 訓練入力: $\{\boldsymbol{x}_i\}_{i=1}^{n} \overset{i.i.d.}{\sim} p_{\mathrm{train}}(\boldsymbol{x})$
  - テスト入力: $\{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'} \overset{i.i.d.}{\sim} p_{\mathrm{test}}(\boldsymbol{x})$
- 真の重要度$w(\boldsymbol{x})$との二乗誤差を最小にするように重要度モデル $w_{\boldsymbol{\alpha}}(\boldsymbol{x})$ を学習: $\min_{\boldsymbol{\alpha}} \left[J(\boldsymbol{\alpha})\right]$

$$J(\boldsymbol{\alpha}) = \frac{1}{2}\int \left(w_{\boldsymbol{\alpha}}(\boldsymbol{x}) - w(\boldsymbol{x})\right)^2 p_{\mathrm{train}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

$$w(\boldsymbol{x}) = \frac{p_{\mathrm{test}}(\boldsymbol{x})}{p_{\mathrm{train}}(\boldsymbol{x})}$$

$$= \frac{1}{2}\int w_{\boldsymbol{\alpha}}(\boldsymbol{x})^2 p_{\mathrm{train}}(\boldsymbol{x})d\boldsymbol{x} - \int w_{\boldsymbol{\alpha}}(\boldsymbol{x})p_{\mathrm{test}}(\boldsymbol{x})d\boldsymbol{x} + C$$

$$\approx \frac{1}{2n}\sum_{i=1}^{n} w_{\boldsymbol{\alpha}}(\boldsymbol{x}_i)^2 - \frac{1}{n'}\sum_{i'=1}^{n'} w_{\boldsymbol{\alpha}}(\boldsymbol{x}'_{i'}) + C$$

# アルゴリズム

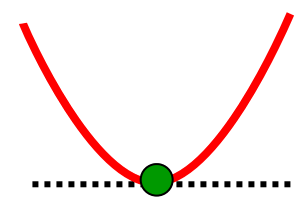
- 重要度モデル: $$w_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{j=1}^{n'} \alpha_j \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}'_j\|^2}{2\sigma^2}\right)$$
- 最適化規準: $$\min_{\boldsymbol{\alpha}} \left[\frac{1}{2}\boldsymbol{\alpha}^\top \widehat{\boldsymbol{G}}\boldsymbol{\alpha} - \widehat{\boldsymbol{h}}^\top \boldsymbol{\alpha} + \underbrace{\frac{\lambda}{2}\boldsymbol{\alpha}^\top \boldsymbol{\alpha}}_{\text{正則化項}}\right]$$

$$\widehat{G}_{j,j'} = \frac{1}{n}\sum_{i=1}^{n} \exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}'_j\|^2}{2\sigma^2}\right) \exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}'_{j'}\|^2}{2\sigma^2}\right)$$

$$\widehat{h}_j = \frac{1}{n'}\sum_{i'=1}^{n'} \exp\left(-\frac{\|\boldsymbol{x}'_{i'}-\boldsymbol{x}'_j\|^2}{2\sigma^2}\right)$$

- 大域的最適解が解析的に計算可能:

$$\widehat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{G}} + \lambda \boldsymbol{I})^{-1}\widehat{\boldsymbol{h}}$$



- $\sigma, \lambda$ は二乗誤差 $J$ に関する交差確認により決定

# 重要度の推定例

- ガウスカーネル重要度モデル

$$w_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{j=1}^{n'} \alpha_j \exp\left( -\frac{\|\boldsymbol{x} - \boldsymbol{x}'_j\|^2}{2\sigma^2} \right)$$

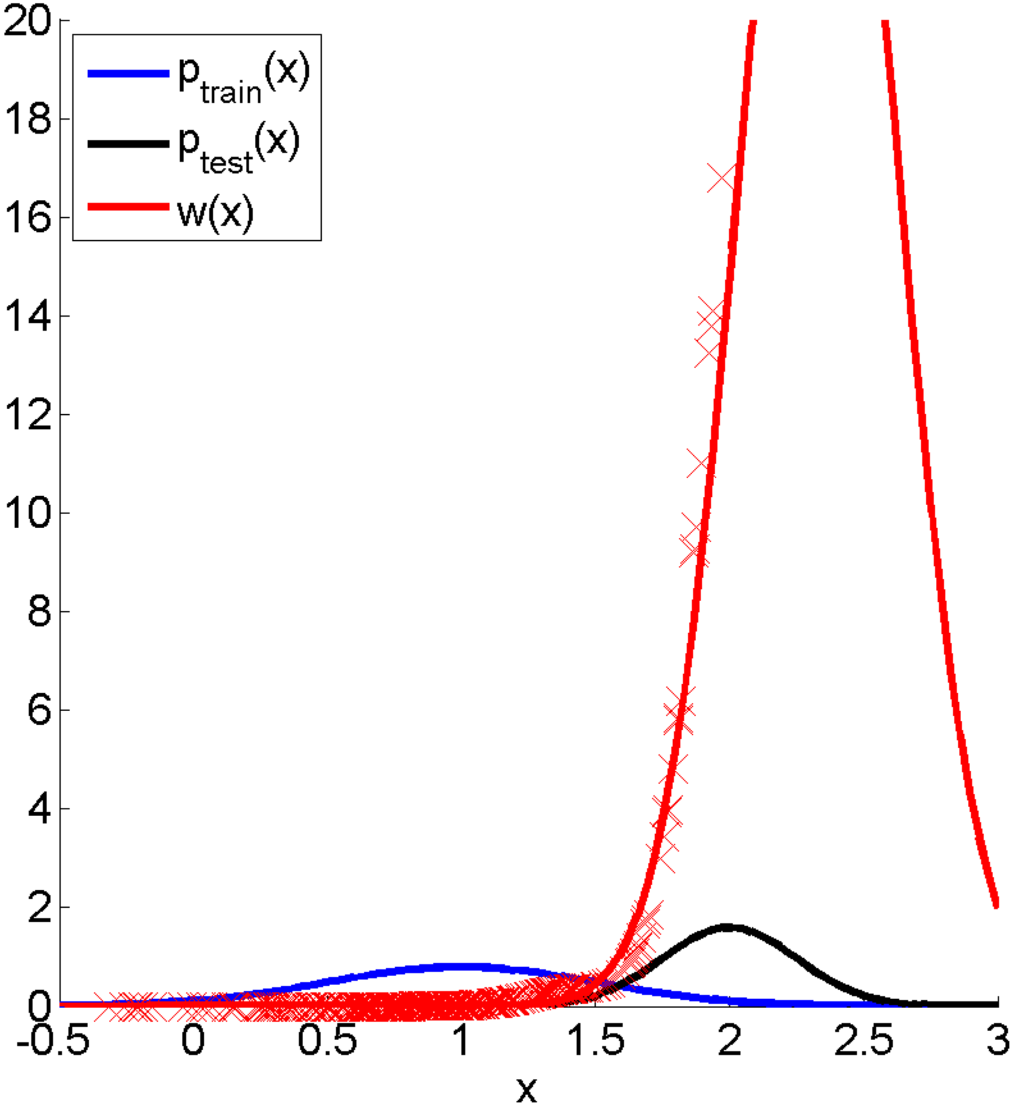
に対する最小二乗重要度推定法

```
clear all; rand('state',0); randn('state',0);
n=100; x=randn(n,1)/4+1; u=randn(n,1)/2;
x2=x.^2; xx=repmat(x2,1,n)+repmat(x2',n,1)-2*x*x';
u2=u.^2; ux=repmat(u2,1,n)+repmat(x2',n,1)-2*u*x';

k=exp(-xx/0.1); r=exp(-ux/0.1);
w=r*((r'*r/n+0.1*eye(n))\(mean(k)'));
figure(1); clf; hold on; plot(u,w,'rx');
```

# 重要度の推定例(続き)





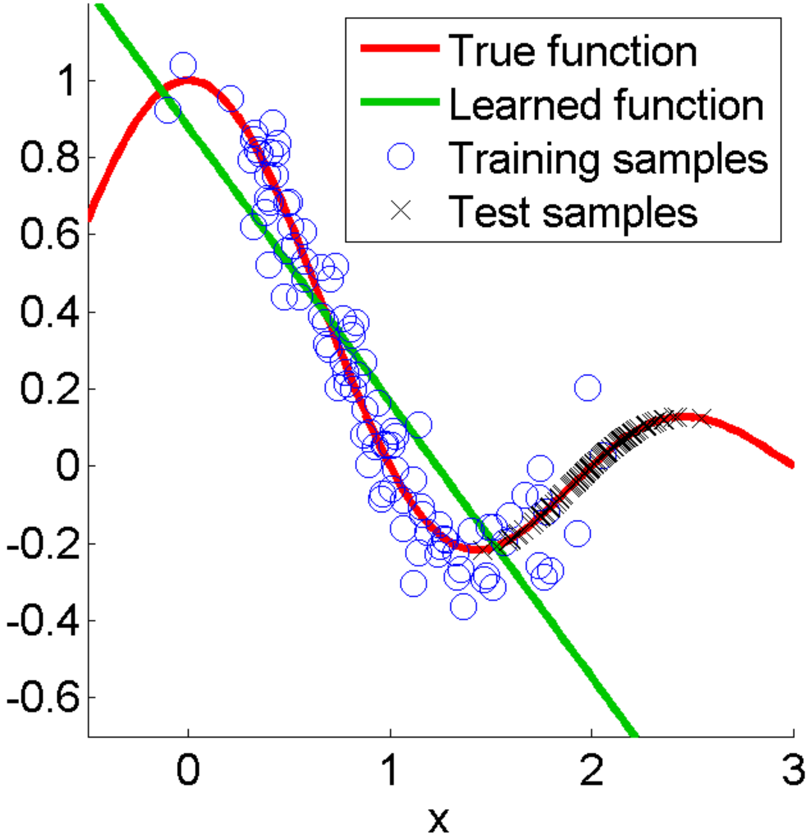

$$w(\boldsymbol{x}) = \frac{p_{\mathrm{test}}(\boldsymbol{x})}{p_{\mathrm{train}}(\boldsymbol{x})}$$

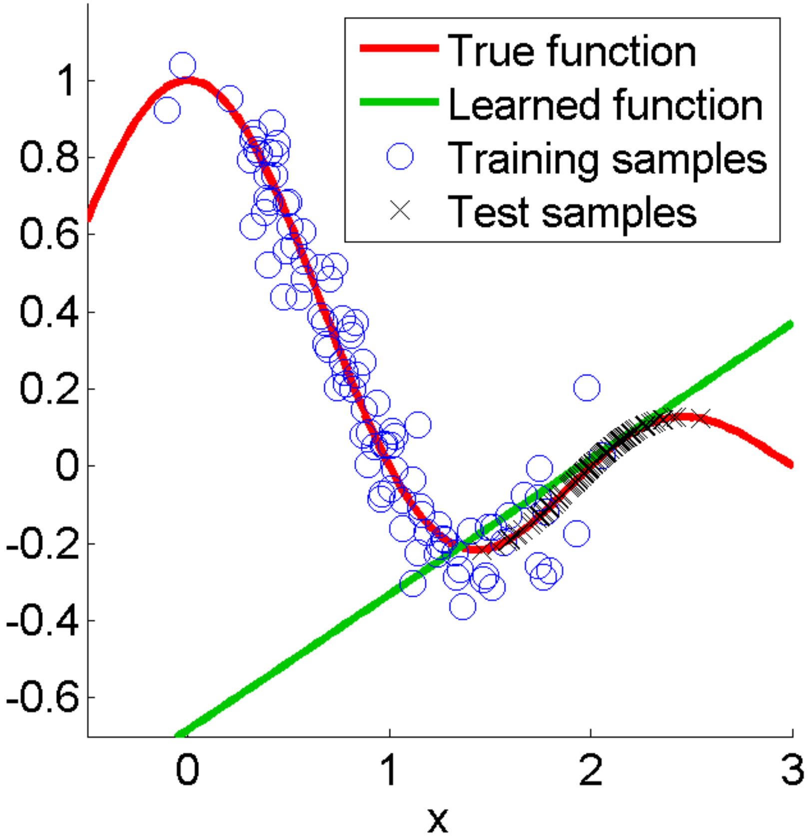
- 重要度が精度良く推定できている

# 重要度重み付き最小二乗法の実行例

$$\min_{\boldsymbol{\theta}} \left[ \sum_{i=1}^{n} \left( f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i \right)^2 \right]$$

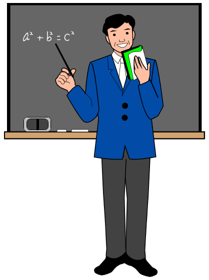
$$\min_{\boldsymbol{\theta}} \left[ \sum_{i=1}^{n} \frac{p_{\mathrm{test}}(\boldsymbol{x}_i)}{p_{\mathrm{train}}(\boldsymbol{x}_i)} \left( f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i \right)^2 \right]$$



# 共変量シフト：まとめ

- 転移学習：
  - 異なる確率分布に従う訓練標本を用いて，テスト標本の出力を精度良く予測したい
- 共変量シフト：
  - 入力の確率分布のみが変化する
- 重要度（確率密度の比）で重みを付けて学習すれば，確率分布の違いを補正できる
- 重要度は，個々の確率密度を推定することなく，直接推定できる



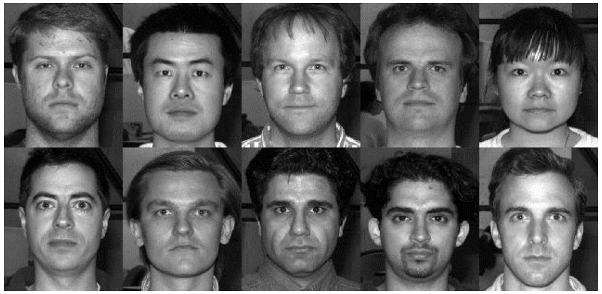
# 講義の流れ



1. 半教師付き学習（16章）
2. 転移学習（18章）
   A) 共変量シフト
   B) クラスバランス変化

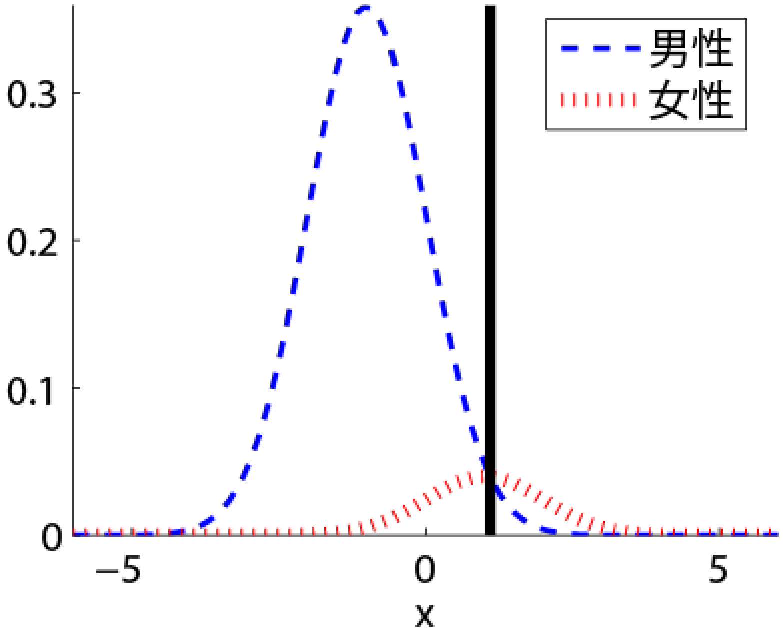
# クラス比変化

- 顔画像からの性別予測
  - 訓練標本：大学には女性が少ない
  - テスト標本：世間での男女比は一対一
- 識別境界がずれる

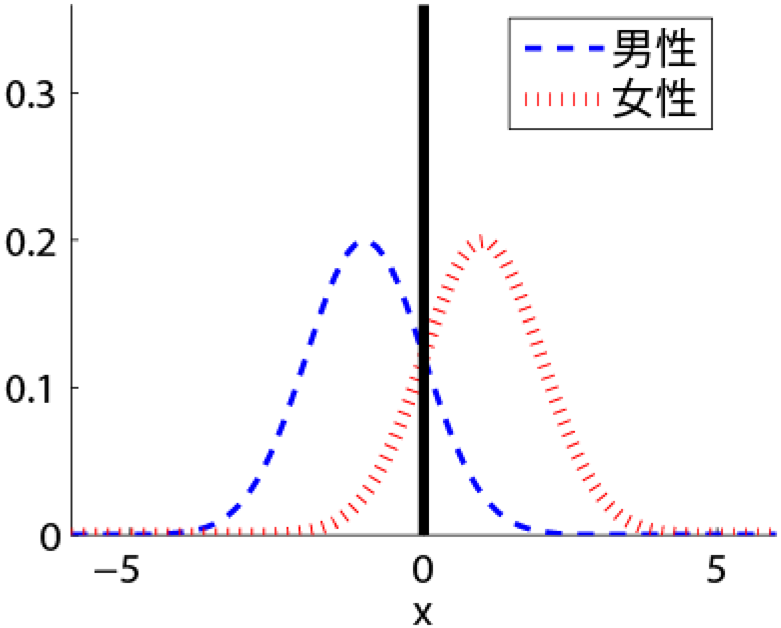


The Yale Face Database B

訓練標本




テスト標本



# クラス比重み付き学習

- テスト標本と比べて，訓練標本中で
  - 比率の小さいクラスの標本には大きな重みを与える
  - 比率の大きいクラスの標本には小さな重みを与える

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \frac{p_{\mathrm{test}}(y_i)}{p_{\mathrm{train}}(y_i)} \left( f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i \right)^2$$

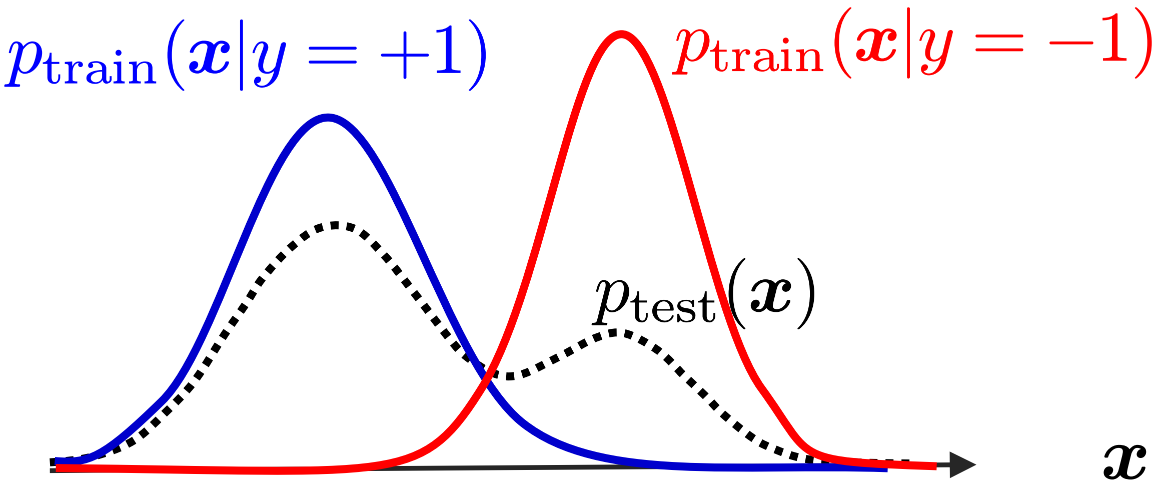
- テスト標本のクラスバランスがわからない場合はどうすればよいか？

# クラス比の推定

- テスト入力の分布 $p_{\mathrm{test}}$ が各クラスの訓練入力の分布の混合分布 $q_\pi$ で表されると仮定すると，真の混合比 $\pi$ は(何らかの)距離尺度を最小化：

$$\min_{\pi\in[0,1]} \mathrm{dist}\left(p_{\mathrm{test}} \middle\| q_\pi\right)$$

$$q_\pi(\boldsymbol{x}) = \pi p_{\mathrm{train}}(\boldsymbol{x}|y=+1) + (1-\pi)p_{\mathrm{train}}(\boldsymbol{x}|y=-1)$$



# 確率分布間の距離

- エネルギー距離(の二乗):

$$D_{\mathrm{E}}^2(p_{\mathrm{test}}, q_\pi) = 2\mathbb{E}_{\boldsymbol{x}' \sim p_{\mathrm{test}}, \boldsymbol{x} \sim q_\pi} \|\boldsymbol{x}' - \boldsymbol{x}\|$$

$$-\mathbb{E}_{\boldsymbol{x}', \widetilde{\boldsymbol{x}}' \sim p_{\mathrm{test}}} \|\boldsymbol{x}' - \widetilde{\boldsymbol{x}}'\| - \mathbb{E}_{\boldsymbol{x}, \widetilde{\boldsymbol{x}} \sim q_\pi} \|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|$$

  - $D_{\mathrm{E}}(p_{\mathrm{test}}, q_\pi) \geq 0$
  - $D_{\mathrm{E}}(p_{\mathrm{test}}, q_\pi) = 0 \iff p_{\mathrm{test}} = q_\pi$

$$q_\pi(\boldsymbol{x}) = \pi p_{\mathrm{train}}(\boldsymbol{x}|y=+1) + (1-\pi) p_{\mathrm{train}}(\boldsymbol{x}|y=-1)$$

- 期待値で定義されるため，標本平均で簡単に近似できる

# エネルギー距離

$$D_{\mathrm{E}}^2(p_{\mathrm{test}}, q_\pi) = 2\mathbb{E}_{\boldsymbol{x}' \sim p_{\mathrm{test}}, \boldsymbol{x} \sim q_\pi} \|\boldsymbol{x}' - \boldsymbol{x}\|$$

$$-\mathbb{E}_{\boldsymbol{x}', \widetilde{\boldsymbol{x}}' \sim p_{\mathrm{test}}} \|\boldsymbol{x}' - \widetilde{\boldsymbol{x}}'\| - \mathbb{E}_{\boldsymbol{x}, \widetilde{\boldsymbol{x}} \sim q_\pi} \|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|$$

$$q_\pi(\boldsymbol{x}) = \pi p_{\mathrm{train}}(\boldsymbol{x}|y=+1) + (1-\pi) p_{\mathrm{train}}(\boldsymbol{x}|y=-1)$$

- エネルギー距離(の二乗)は, $\pi$ の関数として次のように表現できる:

証明は宿題

$$J(\pi) = (2A_{+1,-1} - A_{+1,+1} - A_{-1,-1})\pi^2$$

$$-2(A_{+1,-1} - A_{-1,-1} - b_{+1} + b_{-1})\pi + \mathrm{Const.}$$

$$A_{y,\widetilde{y}} = \mathbb{E}_{\boldsymbol{x} \sim p_{\mathrm{train}}(\boldsymbol{x}|y), \widetilde{\boldsymbol{x}} \sim p_{\mathrm{train}}(\boldsymbol{x}|\widetilde{y})} \|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|$$

$$b_y = \mathbb{E}_{\boldsymbol{x}' \sim p_{\mathrm{test}}, \boldsymbol{x} \sim p_{\mathrm{train}}(\boldsymbol{x}|y)} \|\boldsymbol{x}' - \boldsymbol{x}\|$$

# エネルギー距離の標本近似

$$A_{y,\widetilde{y}} = \mathbb{E}_{\boldsymbol{x} \sim p_{\mathrm{train}}(\boldsymbol{x}|y), \widetilde{\boldsymbol{x}} \sim p_{\mathrm{train}}(\boldsymbol{x}|\widetilde{y})} \|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|$$

$$b_y = \mathbb{E}_{\boldsymbol{x}' \sim p_{\mathrm{test}}, \boldsymbol{x} \sim p_{\mathrm{train}}(\boldsymbol{x}|y)} \|\boldsymbol{x}' - \boldsymbol{x}\|$$

- 期待値を標本平均で近似すれば,

$$\widehat{J}(\pi) = (2\widehat{A}_{+1,-1} - \widehat{A}_{+1,+1} - \widehat{A}_{-1,-1})\pi^2$$

$$-2(\widehat{A}_{+1,-1} - \widehat{A}_{-1,-1} - \widehat{b}_{+1} + \widehat{b}_{-1})\pi + \mathrm{const.}$$

$$\widehat{A}_{y,\widetilde{y}} = \frac{1}{n_y n_{\widetilde{y}}} \sum_{i:y_i=y} \sum_{\widetilde{i}:y_{\widetilde{i}}=\widetilde{y}} \|\boldsymbol{x}_i - \boldsymbol{x}_{\widetilde{i}}\|$$

$$\widehat{b}_y = \frac{1}{n' n_y} \sum_{i'=1}^{n'} \sum_{i:y_i=y} \|\boldsymbol{x}'_{i'} - \boldsymbol{x}_i\|$$

$$\{\boldsymbol{x}_i\}_{i=1}^{n} \overset{\mathrm{i.i.d.}}{\sim} p_{\mathrm{train}}(\boldsymbol{x})$$

$$\{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'} \overset{\mathrm{i.i.d.}}{\sim} p_{\mathrm{test}}(\boldsymbol{x})$$

$n_y$ : クラス $y$ の標本数

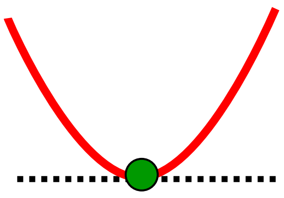
# エネルギー距離に基づくクラス比の推定

$$\widehat{J}(\pi) = (2\widehat{A}_{+1,-1} - \widehat{A}_{+1,+1} - \widehat{A}_{-1,-1})\pi^2$$
$$-2(\widehat{A}_{+1,-1} - \widehat{A}_{-1,-1} - \widehat{b}_{+1} + \widehat{b}_{-1})\pi + \text{const.}$$

- クラス比の推定値

$$\widehat{\pi} = \underset{0 \le \pi \le 1}{\operatorname{argmin}} J(\pi)$$

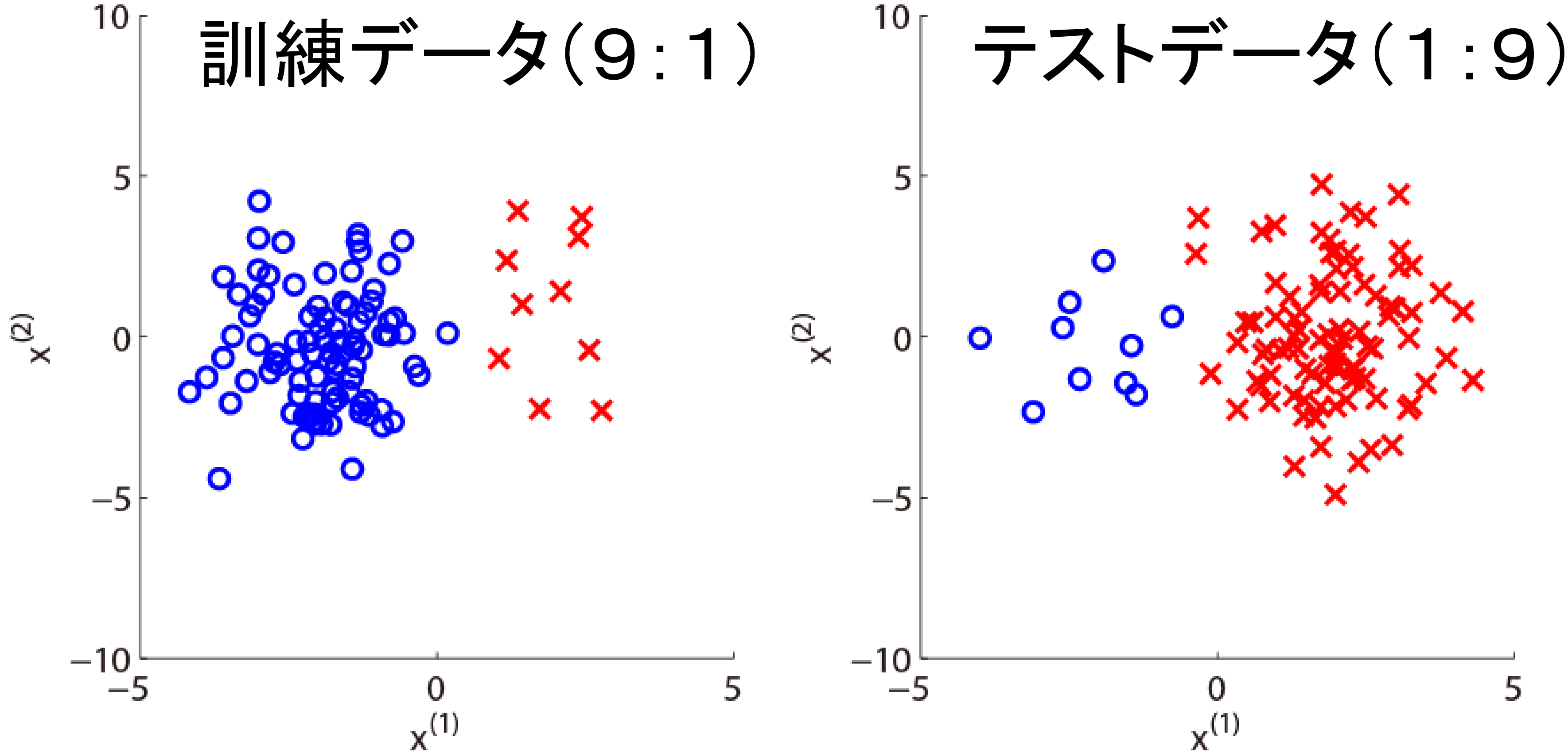
は解析的に求められる:



$$\widetilde{\pi} = \frac{\widehat{A}_{+1,-1} - \widehat{A}_{-1,-1} - \widehat{b}_{+1} + \widehat{b}_{-1}}{2\widehat{A}_{+1,-1} - \widehat{A}_{+1,+1} - \widehat{A}_{-1,-1}}$$

$$\widehat{\pi} = \min(1, \max(0, \widetilde{\pi}))$$

# クラス比の推定例



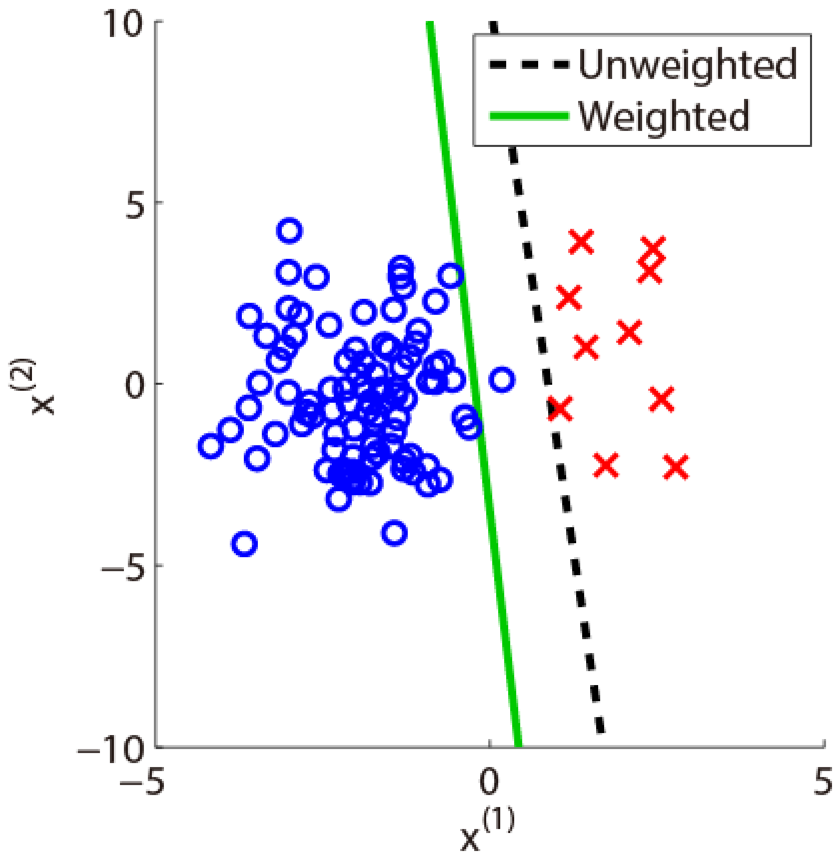

- テストデータのクラスバランス推定結果

(０. １８:０. ８２)

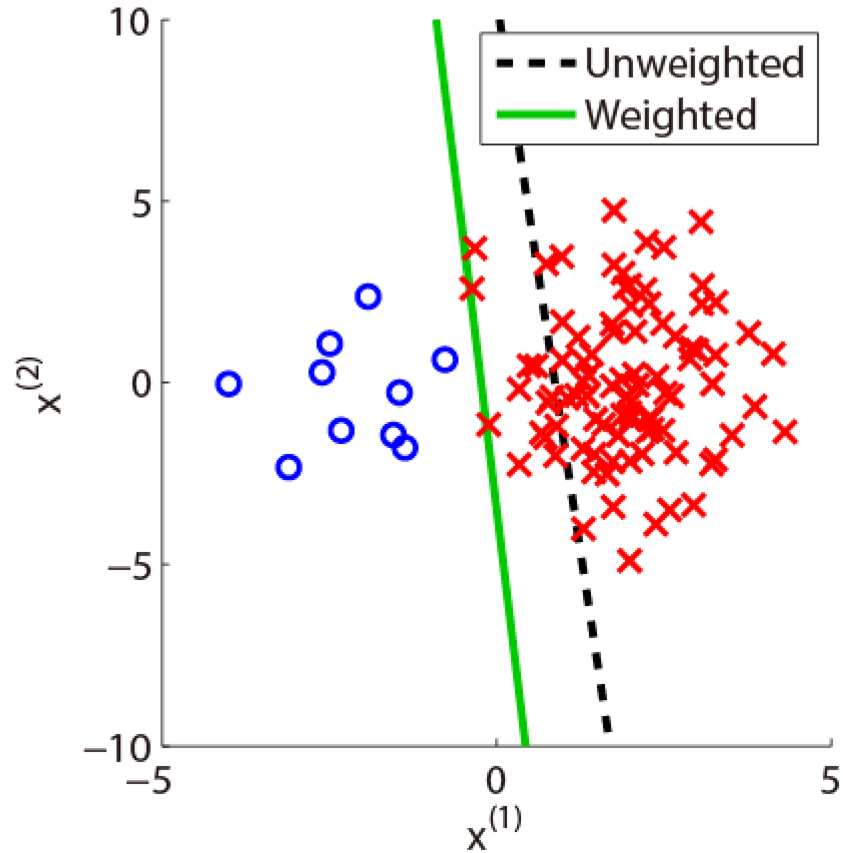
# クラス比重み付き最小二乗法の実行例

- テストデータに対する分類精度が向上している

訓練データ



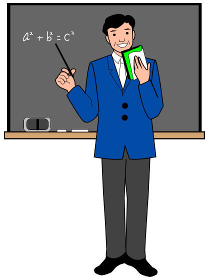

テストデータ



# クラス比変化：まとめ



- 転移学習：
  - 異なる確率分布に従う訓練標本を用いて，テスト標本の出力を精度良く予測したい
- クラス比変化：
  - クラスのバランスのみが変化する
- 重要度（クラス比）で重みを付けて学習すれば確率分布の違いを補正できる
- 重要度は，適当な距離尺度のもとでの適合によって推定できる



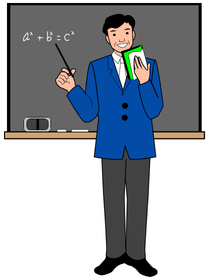
# 講義の流れ



1. 半教師付き学習（16章）
2. 転移学習（18章）

# まとめ

- 半教師付き学習：ラベルなしデータを活用する
  - 同じかたまりに属するデータは同じクラスに属すると仮定
  - ラプラス正則化によりラベルを伝播
- 転移学習：訓練時とテスト時でデータの生成分布が変化
  - 共変量シフト：入力分布が変化
  - クラス比変化：各クラスの標本のバランスが変化
  - 重要度重み付き学習により適応
  - 確率分布を推定せず，重要度を直接推定する



# 次回の予告



- 線形次元削減（13章，17章）

# 宿題1

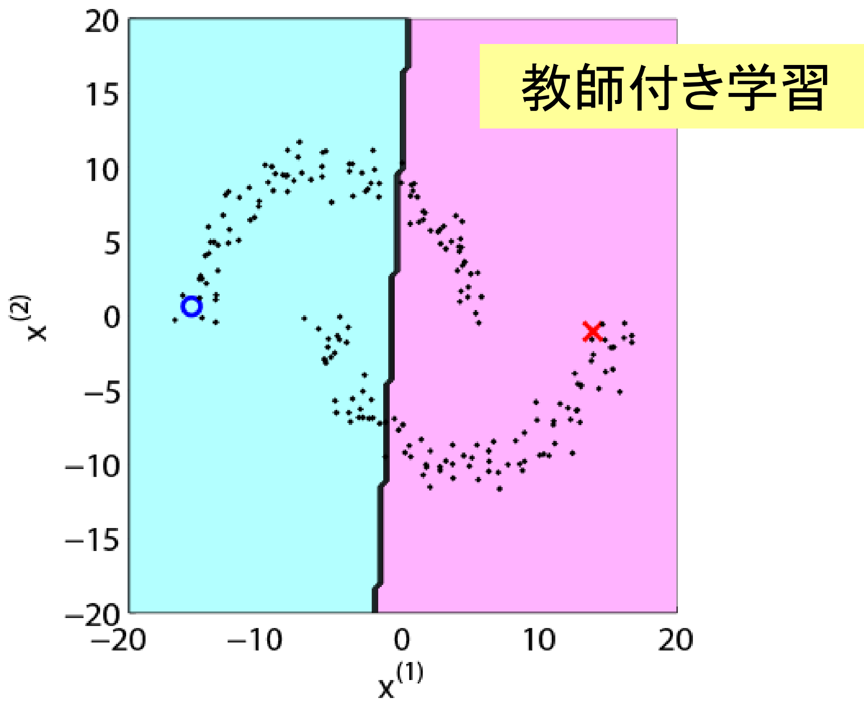


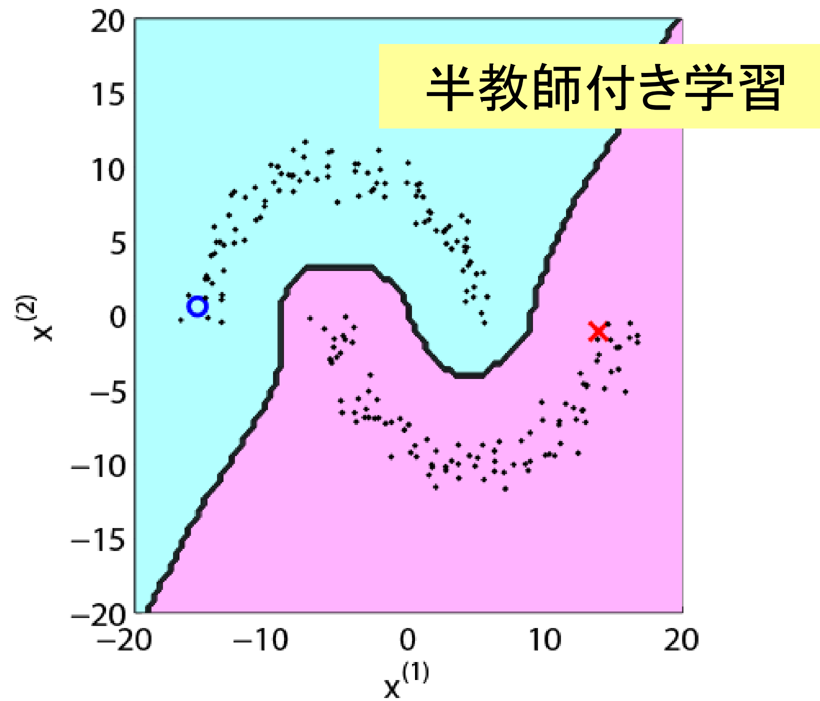
## ■ガウスカーネルモデル

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{n+n'} \theta_j K(\boldsymbol{x}, \boldsymbol{x}_j)$$

$$K(\boldsymbol{x}, \boldsymbol{c}) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{c}\|^2}{2h^2}\right)$$

に対してラプラス正則化最小二乗分類を実装せよ



# 宿題1（続き）



$$\widehat{\boldsymbol{\theta}} = \mathop{\mathrm{argmin}}_{\boldsymbol{\theta}} \left[ \sum_{i=1}^{n} \left( f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i \right)^2 + \lambda \|\boldsymbol{\theta}\|^2 + \nu \sum_{i,i'=1}^{n+n'} W_{i,i'} \left( f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - f_{\boldsymbol{\theta}}(\boldsymbol{x}_{i'}) \right)^2 \right]$$

- 正則化項には例えば$\lambda = \nu = 1$ を用いてよい
- 近傍グラフの重みにはガウスカーネル

$$W_{i,i'} = \exp\left( -\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|^2}{2h^2} \right)$$

を用いてよい

# 宿題1（続き）



```
clear all; rand('state',0); randn('state',0);
n=200; a=linspace(0,pi,n/2);
u=-10*[cos(a)+0.5 cos(a)-0.5]'+randn(n,1);
v=10*[sin(a) -sin(a)]'+randn(n,1);
x=[u v];  % Training input
y=zeros(n,1); y(1)=-1; y(n)=1; % Training output
hh=2*1^2; % Gauss width (2h^2)

t=LRLS(x,y,hh); % Parameter learning

m=100; X=linspace(-20,20,m)'; X2=X.^2;
U=exp(-(repmat(u.^2,1,m)+repmat(X2',n,1)-2*u*X')/hh);
V=exp(-(repmat(v.^2,1,m)+repmat(X2',n,1)-2*v*X')/hh);
figure(1); clf; hold on; axis([-20 20 -20 20]);
colormap([1 0.7 1; 0.7 1 1]);
contourf(X,X,sign(V'*(U.*repmat(t,1,m))));
plot(x(y==1,1),x(y==1,2),'bo');
plot(x(y==-1,1),x(y==-1,2),'rx');
plot(x(y==0,1),x(y==0,2),'k.');
```

LRLS.m
を実装せよ

# 宿題2

$$D_{\mathrm{E}}^2(p_{\mathrm{test}}, q_\pi) = 2\mathbb{E}_{\boldsymbol{x}'\sim p_{\mathrm{test}}, \boldsymbol{x}\sim q_\pi}\|\boldsymbol{x}' - \boldsymbol{x}\|$$
$$-\mathbb{E}_{\boldsymbol{x}', \widetilde{\boldsymbol{x}}'\sim p_{\mathrm{test}}}\|\boldsymbol{x}' - \widetilde{\boldsymbol{x}}'\| - \mathbb{E}_{\boldsymbol{x}, \widetilde{\boldsymbol{x}}\sim q_\pi}\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|$$

$$q_\pi(\boldsymbol{x}) = \pi p_{\mathrm{train}}(\boldsymbol{x}|y=+1) + (1-\pi)p_{\mathrm{train}}(\boldsymbol{x}|y=-1)$$

- $D_{\mathrm{E}}^2(p_{\mathrm{test}}, q_\pi)$ の $\pi$ に関する以下の表現を導け

$$J(\pi) = (2A_{+1,-1} - A_{+1,+1} - A_{-1,-1})\pi^2$$
$$-2(A_{+1,-1} - A_{-1,-1} - b_{+1} + b_{-1})\pi + \mathrm{Const.}$$

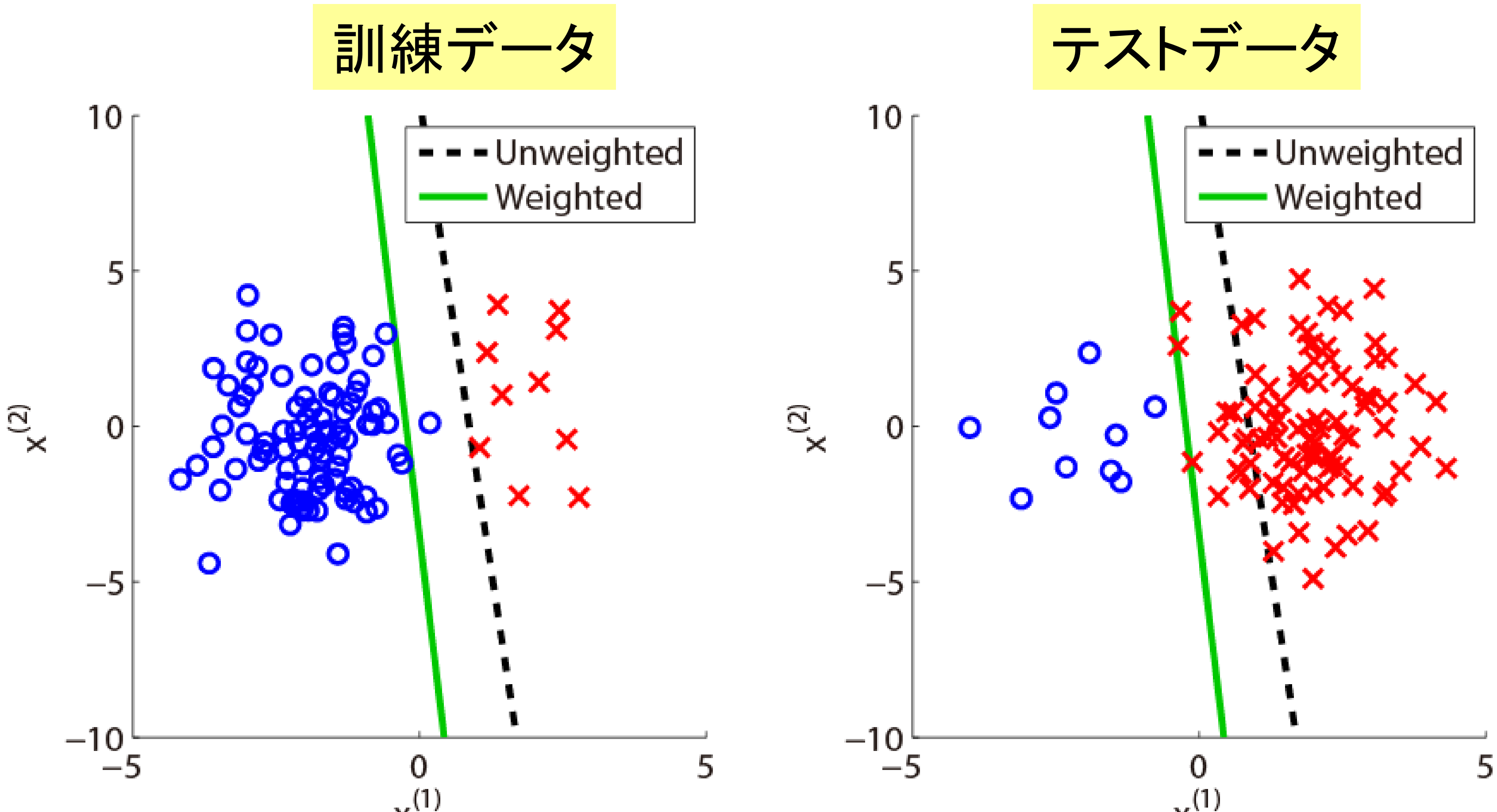
$$A_{y,\widetilde{y}} = \mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{train}}(\boldsymbol{x}|y), \widetilde{\boldsymbol{x}}\sim p_{\mathrm{train}}(\boldsymbol{x}|\widetilde{y})}\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|$$

$$b_y = \mathbb{E}_{\boldsymbol{x}'\sim p_{\mathrm{test}}, \boldsymbol{x}\sim p_{\mathrm{train}}(\boldsymbol{x}|y)}\|\boldsymbol{x}' - \boldsymbol{x}\|$$

- ヒント: $\mathbb{E}_{\widetilde{\boldsymbol{x}}\sim q_\pi}[f(\widetilde{\boldsymbol{x}})] = \pi\mathbb{E}_{\widetilde{\boldsymbol{x}}\sim p_{\mathrm{train}}(\widetilde{\boldsymbol{x}}|+1)}[f(\widetilde{\boldsymbol{x}})]$
$$+(1-\pi)\mathbb{E}_{\widetilde{\boldsymbol{x}}\sim p_{\mathrm{train}}(\widetilde{\boldsymbol{x}}|-1)}[f(\widetilde{\boldsymbol{x}})]$$

# 宿題3

- 線形モデル $f_{\boldsymbol{\theta}}(x) = \boldsymbol{\theta}^\top \boldsymbol{x} + \theta_0$ に対して，クラス比重み付き最小二乗法を実装せよ



# 宿題3（続き）

```
clear all; rand('state',0); randn('state',0);
x=[[randn(90,1)-2; randn(10,1)+2] 2*randn(100,1)];
y=[ones(90,1); 2*ones(10,1)];
X=[[randn(10,1)-2; randn(90,1)+2] 2*randn(100,1)];
% x, y, X: Training input, training output, test input

t=CWLS(x,y,X); % Parameter learning

Y=[ones(10,1); 2*ones(90,1)]; % Training output
figure(1); clf; hold on
plot([-5 5],-(t(3)+[-5 5]*t(1))/t(2),'g-');
plot(X(Y==1,1),X(Y==1,2),'bo');
plot(X(Y==2,1),X(Y==2,2),'rx');
legend('Weighted');
axis([-5 5 -10 10])
```

CWLS.mを実装せよ