(83342) パターン認識特論

張　潮

zhang@u-fukui.ac.jp

0776-27-8477

本日の内容

第1回　パターン認識のための線形代数
第2回　パターン認識のための確率論
第3回　凸最適化の概念
第4回　線形回帰とLMSアルゴリズム
第5回　分類とロジスティック回帰
第6回　一般化線形モデル
第7回　ガウシアン判別分析と単純ベイズ分類器
第8回　サポートベクターマシン(1)
第9回　サポートベクターマシン(2)
第10回　正規化とモデル選択
第11回　k平均法
第12回　混合ガウスモデルとEMアルゴリズム
第13回　因子分析
第14回　主成分分析
第15回　未定

今までの分類アルゴリズムの考え方

- 分類問題
  - 二項分類　→　ロジスティクス回帰
  - 多項分類　→　softmax回帰
- 回帰問題と見なす
  - 最尤推定によってθを最適化 ➡ $p(y|x;\theta)$
- 決定境界で分類
  - $p(y|x;\theta)$ >= 0.5 → y:=1
  - $p(y|x;\theta)$ < 0.5 ← y:=0

判別学習アルゴリズム
discriminative learning algorithms

新しい考え方

- 逆の考え方
  - p(*y*|*x*)の代わりにp(*x*|*y*)をモデリングしましょう
  - 例：ベイズの定理を使えばp(*y*|*x*)が求まる

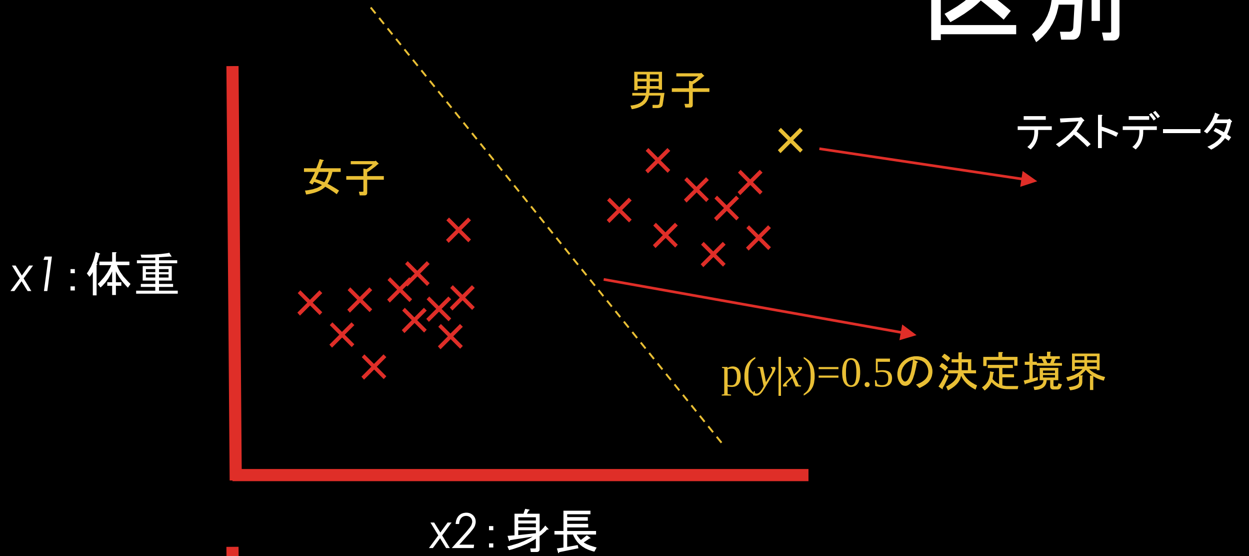
$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}.$$

生成学習アルゴリズム
generative learning algorithms

区別

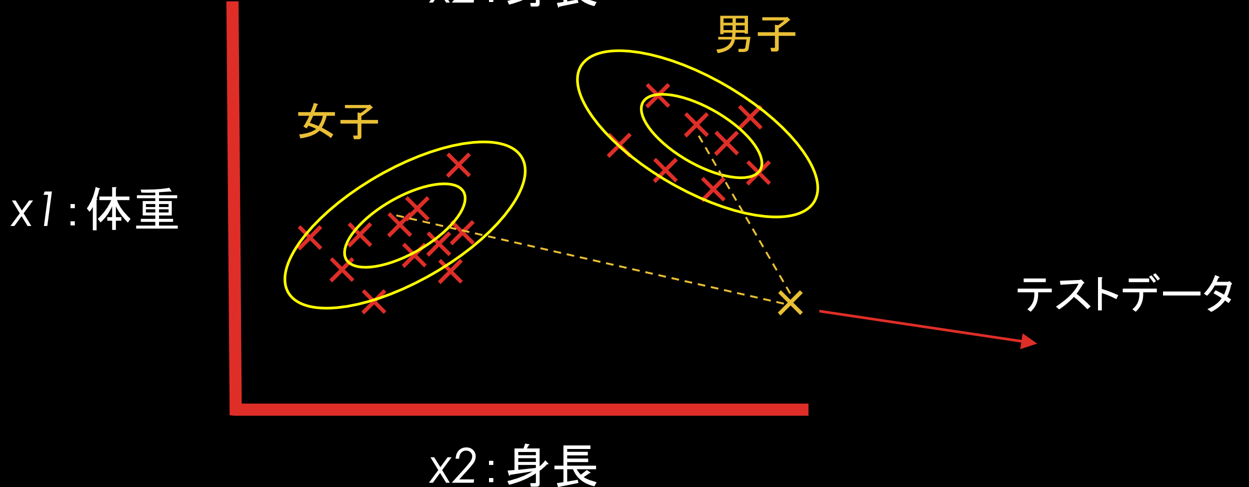
●判別モデル

$$p(y|x)$$



●生成モデル

$$p(x|y)$$



ガウシアン判別分析 Gaussian discriminant analysis (GDA)

- **生成モデル**のアルゴリズム
- 多変量ガウス分布でトレーニングデータの分布をモデリング
- 多変量ガウス分布

$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right).$$

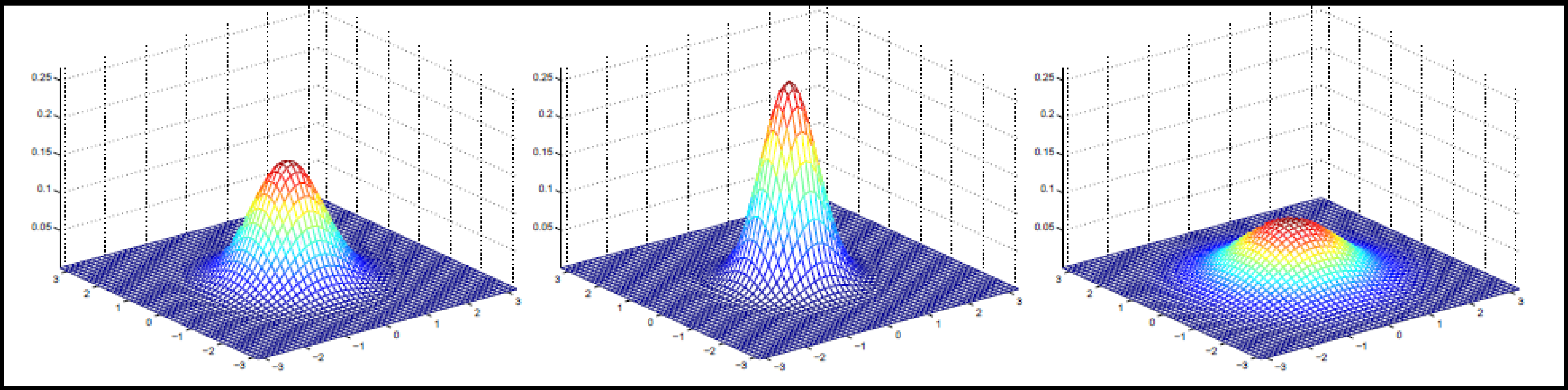
$$\mathcal{N}(\mu,\Sigma)$$

$\mu \in \mathbb{R}^n$: 平均 (mean)

$\Sigma \in \mathbb{R}^{n\times n}$: 共分散行列 (covariance matrix)

ガウシアン判別分析 Gaussian discriminant analysis (GDA)

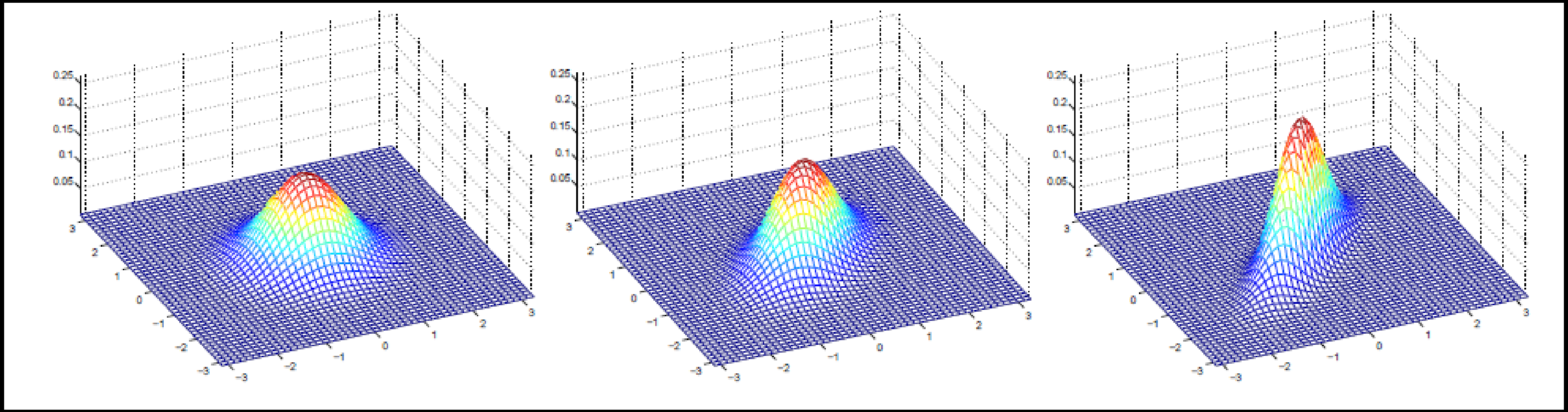
- 共分散行列は単位行列の定数倍のとき



$\Sigma = I$ $\Sigma = 0.6I$ $\Sigma = 2I$

ガウシアン判別分析 Gaussian discriminant analysis (GDA)

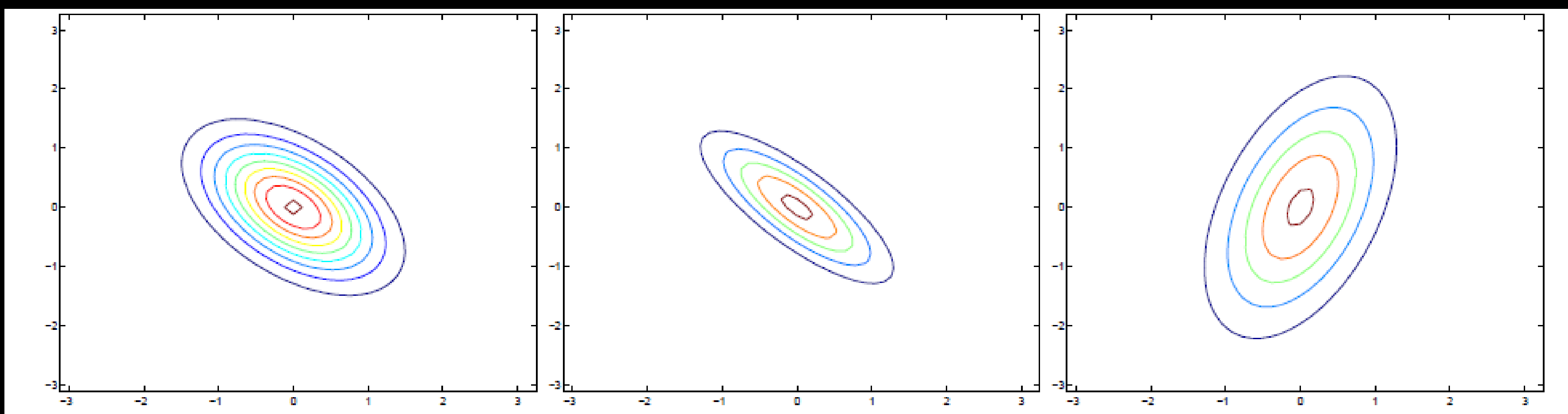
- 共分散行列は単位行列でないとき



$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \quad \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}; \quad .\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

ガウシアン判別分析 Gaussian discriminant analysis (GDA)

- 等高線で2次元の多変量ガウス分布を描画



$$\Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix};\quad \Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix};\quad .\Sigma = \begin{bmatrix} 3 & 0.8 \\ 0.8 & 1 \end{bmatrix}.$$

ガウシアン判別分析 Gaussian discriminant analysis (GDA)

- 二項分類問題
  - 考え方を式で表すと

$$
\begin{aligned}
y &\sim \mathrm{Bernoulli}(\phi) \\
x|y=0 &\sim \mathcal{N}(\mu_0, \Sigma) \\
x|y=1 &\sim \mathcal{N}(\mu_1, \Sigma)
\end{aligned}
$$

$$
\begin{aligned}
p(y) &= \phi^y (1-\phi)^{1-y} \\
p(x|y=0) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^T \Sigma^{-1} (x-\mu_0)\right) \\
p(x|y=1) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1} (x-\mu_1)\right)
\end{aligned}
$$

分布が違うため，平均パラメータが異なる

簡単のため，同じ設定

ガウシアン判別分析 Gaussian discriminant analysis (GDA)

- ガウシアン判別分析における最尤推定
- 対数尤度関数

$$\begin{aligned}\ell(\phi,\mu_0,\mu_1,\Sigma) &= \log\prod_{i=1}^{m} p(x^{(i)},y^{(i)};\phi,\mu_0,\mu_1,\Sigma)\\ &= \log\prod_{i=1}^{m} p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma)p(y^{(i)};\phi).\end{aligned}$$

ガウシアン判別分析 Gaussian discriminant analysis (GDA)

- 対数尤度関数を最大化する

$$\ell(\theta) = \sum_{i=1}^{m} \log p(y^{(i)}|x^{(i)};\theta)$$

同時確率

判別モデル（線形回帰）においての尤度関数

$$\begin{aligned}\ell(\phi,\mu_0,\mu_1,\Sigma) &= \log \prod_{i=1}^{m} p(x^{(i)},y^{(i)};\phi,\mu_0,\mu_1,\Sigma) \\ &= \log \prod_{i=1}^{m} p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma)p(y^{(i)};\phi).\end{aligned}$$

条件付き確率の定義

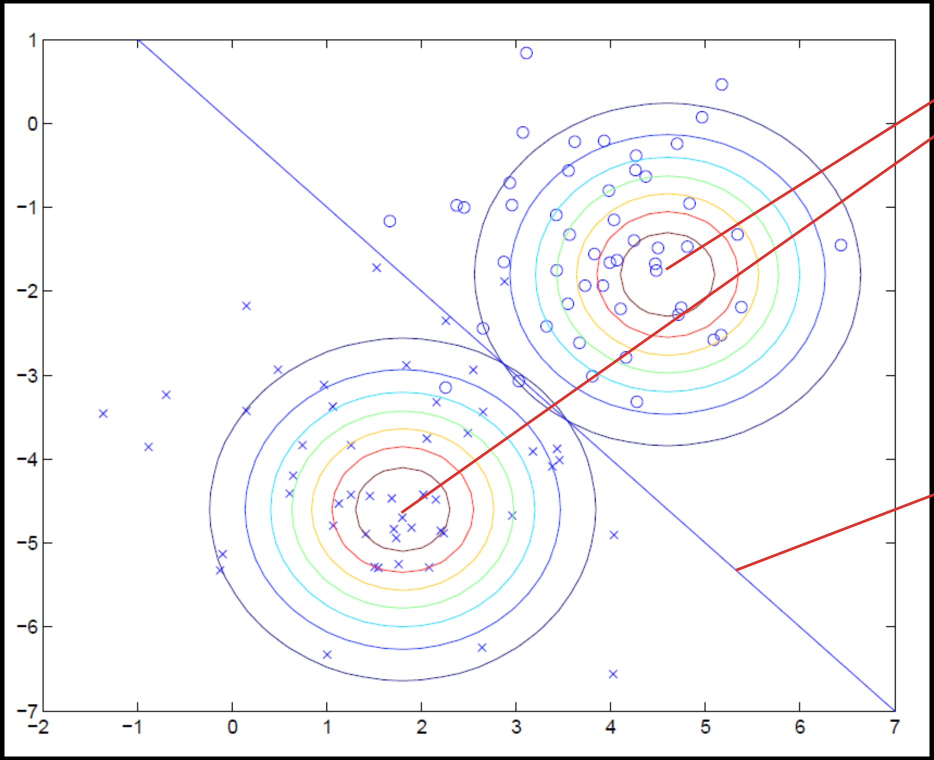
生成モデル（GDA）においての尤度関数

ガウシアン判別分析 Gaussian discriminant analysis (GDA)

- 最適なパラメータが求まる

$$
\begin{aligned}
\phi &= \frac{1}{m}\sum_{i=1}^{m} 1\{y^{(i)} = 1\} \\
\mu_0 &= \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}} \\
\mu_1 &= \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}} \\
\Sigma &= \frac{1}{m}\sum_{i=1}^{m} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T.
\end{aligned}
$$

結果の可視化



生成モデルで
求まった$p(x|y)$

判別モデルで
求まった決定境界

$$p(y=1|x)=0.5$$

判別モデルvs.生成モデル

- トレーニングデータの分布が仮定した分布に近づくほど、生成モデルがより有効と言える
  - non-Gaussian dataにGDAを適用しても効果が低い
- トレーニングデータの分布が仮定した分布に近いと分かれば、生成モデルは判別モデルに比べて、より少ないデータで学習できる
- サンプル数が多いかつサンプルの分布はガウスではない場合
  - ロジスティクス回帰がGDAより有効

単純ベイズ分類器 Naive Bayes

- 今までの分類問題においての特徴量ベクトルx
  - 連続値

    ↓

  - 離散値
- **迷惑メールの分類問題で考えよう**
  - 一通のメールを特徴量ベクトルで表現

$$x = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \left. \begin{array}{l} \text{a} \\ \text{aardvark} \\ \text{aardwolf} \\ \vdots \\ \text{buy} \\ \vdots \\ \text{zygmurgy} \end{array} \right\} \text{用語数（vocabulary）}$$

単純ベイズ分類器 Naive Bayes

- 生成モデル作りましょう！
  - p(x|y)をモデリングする
- 用語数：50,000
  - $x \in \{0,1\}^{50000}$
  - 50,000次元の多変量ガウス分布を作る？→現実的ではない

解決方法

強い仮説を設ける

条件付き確率の定義

各特徴量の次元は条件付き独立である

$$
\begin{aligned}
&p(x_1,\ldots,x_{50000}|y)\\
&\quad= p(x_1|y)p(x_2|y,x_1)p(x_3|y,x_1,x_2)\cdots p(x_{50000}|y,x_1,\ldots,x_{49999})\\
&\quad= p(x_1|y)p(x_2|y)p(x_3|y)\cdots p(x_{50000}|y)
\end{aligned}
$$

単純ベイズ分類器 Naive Bayes

- つまり，p(x|y)を下記通りの計算で近似する
  - （本当は、単語同士が独立していないはず）
  - 例：loveが出てきたら、youが現れる確率が高くなるはず…

$$
\begin{aligned}
&p(x_1, \ldots, x_{50000}|y) \\
&\quad = \quad p(x_1|y)p(x_2|y, x_1)p(x_3|y, x_1, x_2) \cdots p(x_{50000}|y, x_1, \ldots, x_{49999}) \\
&\quad = \quad p(x_1|y)p(x_2|y)p(x_3|y) \cdots p(x_{50000}|y) \\
&\quad = \quad \prod_{i=1}^{n} p(x_i|y)
\end{aligned}
$$

ベルヌーイ分布

単純ベイズ分類器 Naive Bayes

- 最適化したいパラメータ
  - 50,000個 : $\phi_{i|y=1} = p(x_i = 1|y = 1)$
  - 50,000個 : $\phi_{i|y=0} = p(x_i = 1|y = 0)$
  - 1個 : $\phi_y = p(y = 1)$

- 同時分布を利用して尤度関数を作る

$$\mathcal{L}(\phi_y, \phi_{i|y=0}, \phi_{i|y=1}) = \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}).$$

単純ベイズ分類器 Naive Bayes

- 尤度関数を最大化することにより最適解を求める

$$\phi_{j|y=1} = \frac{\sum_{i=1}^{m} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 1\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}$$

y=1の迷惑メールにおいてのある単語x_jの出現頻度

$$\phi_{j|y=0} = \frac{\sum_{i=1}^{m} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 0\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}}$$

y=0の迷惑メールにおいてのある単語x_jの出現頻度

$$\phi_y = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}{m}$$

y=1の迷惑メールの出現頻度

単純ベイズ分類器
Naive Bayes

● 予測

$$p(y=1|x) = \frac{p(x|y=1)p(y=1)}{p(x)} = \frac{\left(\prod_{i=1}^{n} p(x_i|y=1)\right) p(y=1)}{\left(\prod_{i=1}^{n} p(x_i|y=1)\right) p(y=1) + \left(\prod_{i=1}^{n} p(x_i|y=0)\right) p(y=0)},$$

● 多項分類の場合 $x_i \in \{1, 2, \ldots, k_i\}$

$$\begin{aligned} p(x_1, \ldots, x_{50000}|y) &= p(x_1|y)p(x_2|y,x_1)p(x_3|y,x_1,x_2)\cdots p(x_{50000}|y,x_1,\ldots,x_{49999}) \\ &= p(x_1|y)p(x_2|y)p(x_3|y)\cdots p(x_{50000}|y) \\ &= \prod_{i=1}^{n} p(x_i|y) \end{aligned}$$

ベルヌーイ分布の代わりにN=1の多項分布で考えればいい

単純ベイズ分類器 Naive Bayes

- 離散値と連続値が混ざっているような特徴量の場合
  - 離散化を行う

Living area (sq. feet)	< 400	400-800	800-1200	1200-1600	>1600
x_i	1	2	3	4	5

- 広く使われているアルゴリズム
- 多くの場合、GDAより性能がいい

単純ベイズは生成モデル！（ベイズの定理を使っているため、判別モデルに見えるのですが…）

単純ベイズ分類器の改良

- 問題点

$$\phi_{j|y=1} = \frac{\sum_{i=1}^{m} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 1\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}$$

→ y=1の迷惑メールにおいてのある単語x_jの出現頻度

- 例えば、学習データにj番目の単語：hacker が一度も現れなかったら

$$\phi_{j|y=1} = \frac{\sum_{i=1}^{m} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 1\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}$$

→ 0 （まずい！）

$$\begin{aligned} p(y=1|x) &= \frac{\prod_{i=1}^{n} p(x_i|y=1)p(y=1)}{\prod_{i=1}^{n} p(x_i|y=1)p(y=1) + \prod_{i=1}^{n} p(x_i|y=0)p(y=0)} \\ &= \frac{0}{0}. \end{aligned}$$

→ どう予測すれば分からなくなる

単純ベイズ分類器の改良

- 教師あり学習の弱み
  - 見たことないことは予測できない
- Laplace smoothing

$$\phi_j = \frac{\sum_{i=1}^{m} 1\{z^{(i)} = j\}}{m}.$$

⬇

$$\phi_j = \frac{\sum_{i=1}^{m} 1\{z^{(i)} = j\} + 1}{m + k}.$$

☺

Φ_jが0にならない
かつ
$\sum_{j=1}^{k} \phi_j = 1$ が不変

単純ベイズ分類器の改良

- The multi-variate Bernoulli event model

$$x = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \begin{array}{l} \text{a} \\ \text{aardvark} \\ \text{aardwolf} \\ \vdots \\ \text{buy} \\ \vdots \\ \text{zygmurgy} \end{array}$$

- multinomial event model

$$x = \begin{bmatrix} 1 \\ \\ 500 \\ \vdots \\ 1500 \\ \vdots \\ \\ \end{bmatrix} \begin{array}{l} \text{a} \\ \\ \text{boy} \\ \vdots \\ \text{buy} \\ \vdots \\ \\ \end{array}$$

次元数が減る！
（メールによって次元数が違う）

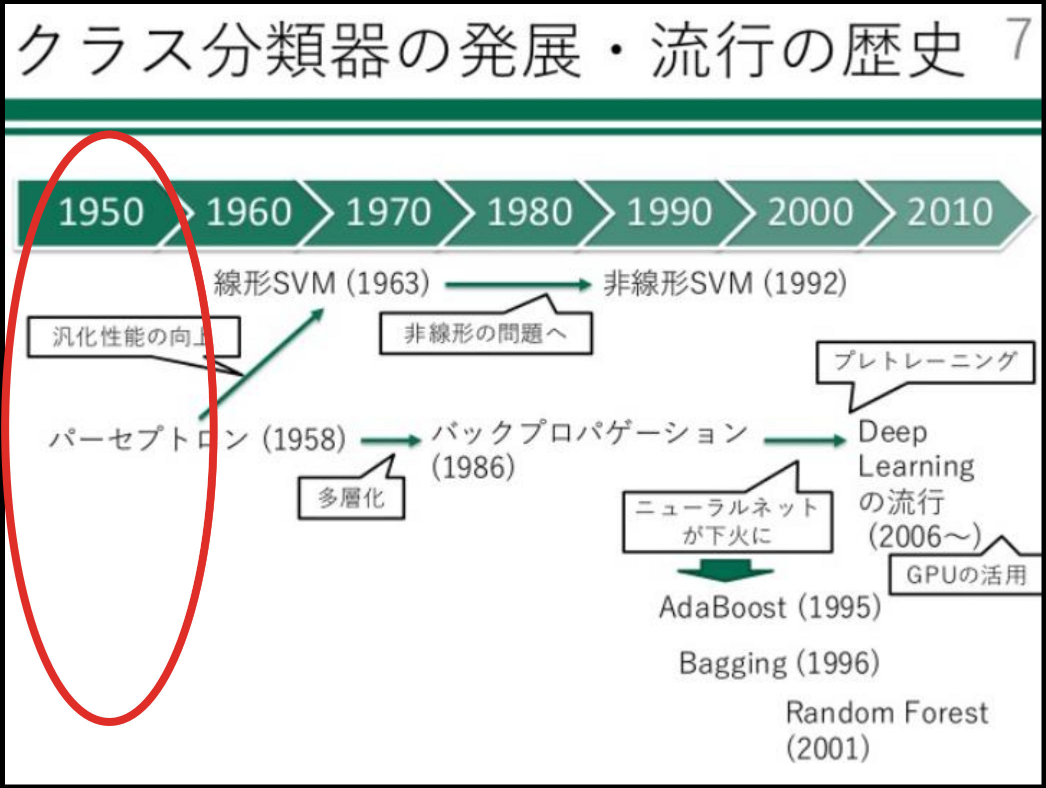
$$\begin{aligned} \phi_{k|y=1} &= \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} 1\{x_j^{(i)} = k \wedge y^{(i)} = 1\} + 1}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\} n_i + |V|} \\ \phi_{k|y=0} &= \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} 1\{x_j^{(i)} = k \wedge y^{(i)} = 0\} + 1}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\} n_i + |V|}. \end{aligned}$$

First thing to try!

余談

- https://image.slidesharecdn.com/tutorialpublic-160607132749/95/python-svmdeep-learning-8-638.jpg?cb=1465372124

分類問題に関して
今まで紹介した内容
←←← ←←← ←←←



次回は

第1回　パターン認識のための線形代数
第2回　パターン認識のための確率論
第3回　凸最適化の概念
第4回　線形回帰とLMSアルゴリズム
第5回　分類とロジスティック回帰
第6回　一般化線形モデル
第7回　ガウシアン判別分析と単純ベイズ分類器
第8回　サポートベクターマシン（1）
第9回　サポートベクターマシン（2）
第10回　正規化とモデル選択
第11回　k平均法
第12回　混合ガウスモデルとEMアルゴリズム
第13回　因子分析
第14回　主成分分析

第15回　未定