

クレジット:

UTokyo Online Education データマイニング入門 2018 森 純一郎

ライセンス:

利用者は、本講義資料を、教育的な目的に限ってページ単位で利用することができます。特に記載のない限り、本講義資料はページ単位でクリエイティブ・コモンズ 表示-非営利-改変禁止 ライセンスの下に提供されています。

<http://creativecommons.org/licenses/by-nc-nd/4.0/>

本講義資料内には、東京大学が第三者より許諾を得て利用している画像等や、各種ライセンスによって提供されている画像等が含まれています。個々の画像等を本講義資料から切り離して利用することはできません。個々の画像等の利用については、それぞれの権利者の定めるところに従ってください。



データマイニング入門 第10回

2018年度

学習目標

- ロジスティック回帰について理解する
 - 2クラス分類について理解する
 - シグモイド関数について理解する
 - 決定境界について理解する
 - コスト関数（交差エントロピー）の最小化について理解する
 - パラメータの推定について理解する
- 多クラス分類について理解する
 - ソフトマックス関数

線形回帰 変数変換と非線形性

入力変数（特徴量）を変換することで非線形性を導入

* モデルは線形回帰

- 対数変換

- 1変数入力

$$h(x) = \theta_0 + \theta_1 \log x$$

- 交差項

- 2変数入力

$$h(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2$$

- 多項式

- 1変数入力の2次の多項式

$$h(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$

- 2変数入力の2次の多項式

$$h(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2 + \theta_4 x_1^2 + \theta_5 x_2^2$$

基底関数

変数変換のより一般的な形を例として1変数入力の3次の多項式で考える

$$h(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$

ここで以下のような非線形関数 $\phi(x)$ （これを基底関数と呼ぶ）を考える

$$\phi(x) = (1, x, x^2, x^3) = (1, \phi_1(x), \phi_2(x), \phi_3(x))$$

$\phi(x)$ は入力 x を（より回帰・識別しやすい）
高次の特徴量空間へ写像している

この時、入力 X は以下のように表すことができる

$$\Phi(x) = (\phi(x^{(1)}), \dots, \phi(x^{(i)}), \dots, \phi(x^{(m)}))^T$$

この正規方程式の解は

$$\theta = (\Phi(x)^T \Phi(x))^{-1} \Phi(x)^T y$$

過学習

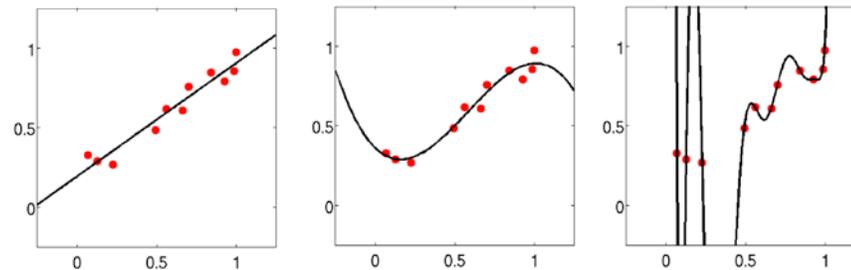
1変数入力の多項式による線形回帰を考えてみる

$$h(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \dots + \theta_n x^n$$

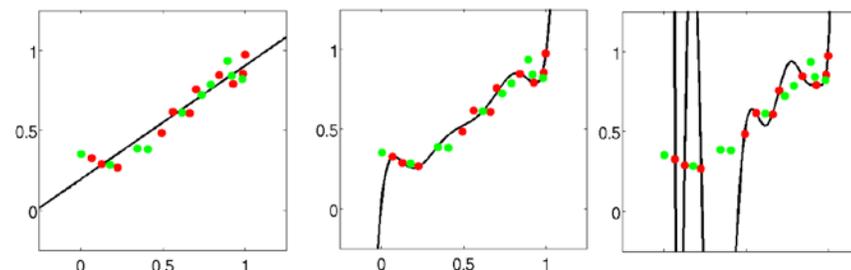
多項式の次数が過度に大きく（仮説関数が複雑に）なると、訓練データに適合しすぎてしまう（訓練データの誤差は小さくなる）

一方、テストデータ（未知のデータ）での予測がうまくいかなくなる

訓練データ（赤点）
への適合



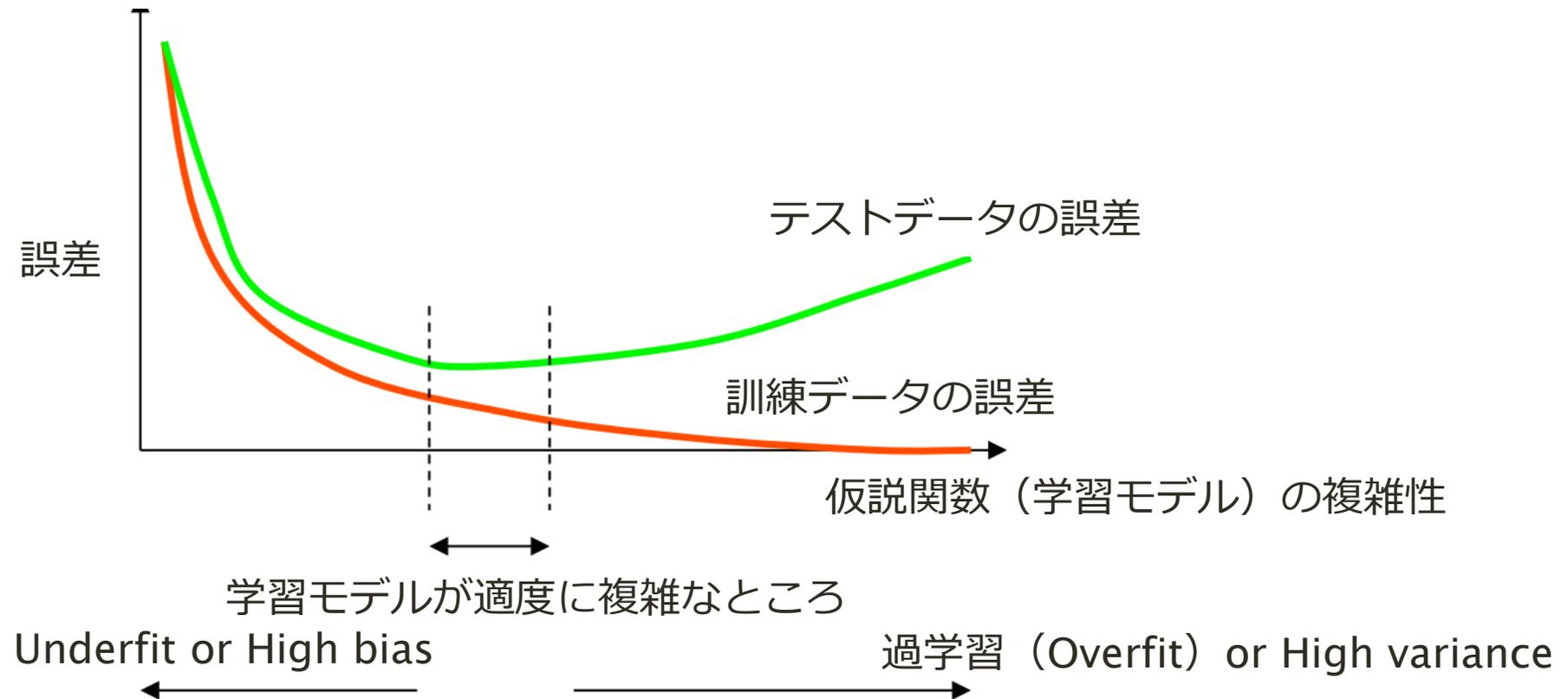
テストデータ（緑点）
での予測



多項式の次数が過度に大きく
仮説関数が複雑になると
訓練データでの誤差は小さくなるが
テストデータでの誤差が大きくなる

過学習

多項式の次数が大きい時のように仮説関数（学習モデル）が複雑なほど過学習が起こりやすい
訓練データだけでなくテストデータでも十分に汎化されているような学習モデルを構築したい



正則化

過学習（パラメータの肥大化）が起こったらペナルティを課すような項（正則化項または罰則項）をコスト関数に追加する

- ・ λ はハイパーパラメータ（大きいほど過学習を抑える）

$$J(\theta) = \frac{1}{2m} (\sum_{i=1}^m (h(x^{(i)}) - y^{(i)})^2 + \lambda \theta^T \theta)$$

* θ_0 は含めない

正則化を用いた最急勾配法によるパラメータ更新

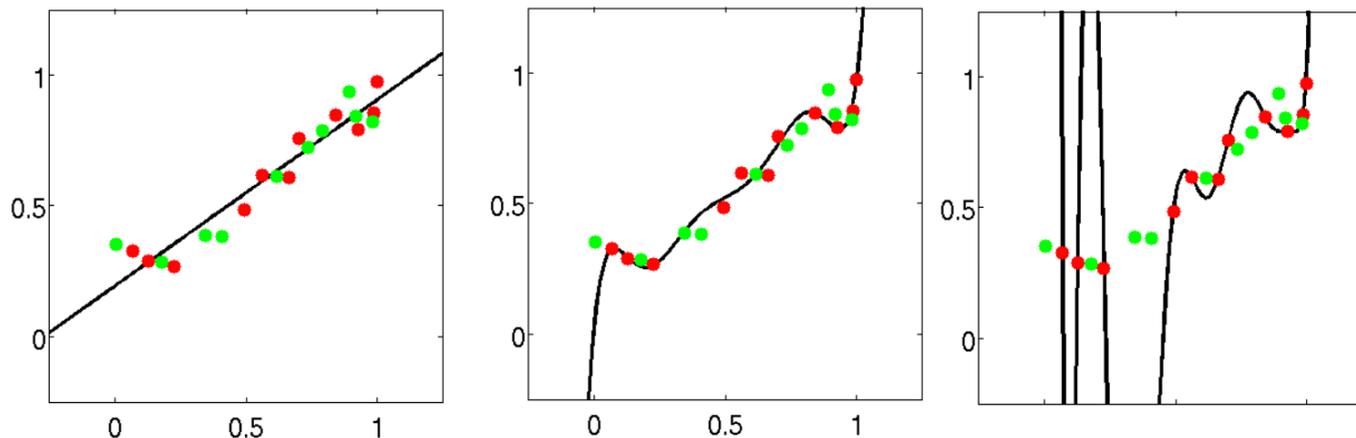
- ・ 正則化項のペナルティの効果によりパラメータが徐々に小さくなり肥大化が抑えられる
- ・ バイアス項に対するパラメータ θ_0 へはペナルティを与えないようにする

$$\theta_j := \theta_j - \alpha \frac{\partial J(\theta)}{\partial \theta_j} = \theta_j (1 - \alpha \frac{\lambda}{m}) - \alpha \sum_{i=1}^m ((h(x^{(i)}) - y^{(i)}) x_j^{(i)}) / m$$

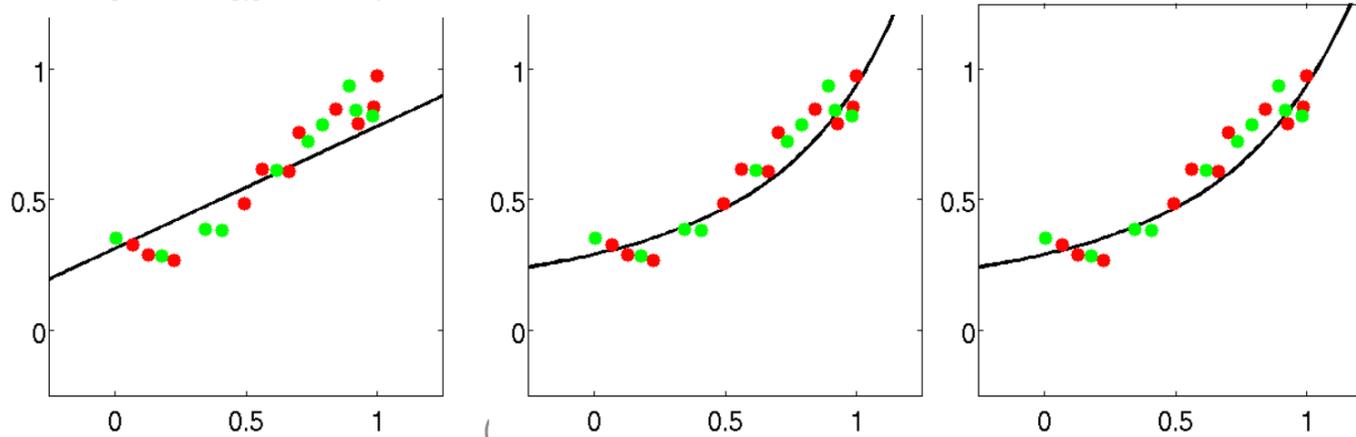
$$\theta_0 := \theta_0 - \alpha \sum_{i=1}^m (h(x^{(i)}) - y^{(i)}) / m$$

正則化

正則化なし



正則化あり
($\lambda=1$)



学習モデルが複雑でも
正則化の効果により
訓練データでの過学習
が抑えられている

正則化 正規方程式による解釈

コスト関数を最小化する（コスト関数の勾配が0になる）ようなパラメータの解析解
正則化なし（正規方程式）

$$\theta = (X^T X)^{-1} X^T y$$

正則化あり

$$\theta = (X^T X + \lambda I)^{-1} X^T y$$

λ は対角要素のみが1であるような行列（ただし x_0 に対応する対角要素は0）
データ数 \ll 特徴量数（学習モデルが複雑）の時、正規方程式の解は不定
 $X^T X$ の対角成分に正の定数を加えて $X^T X$ を正則にすることで解を求められるようにしている

線形回帰

復習：最尤推定による解釈

訓練データセットの出力 $y^{(i)}$ は平均 $h(x^{(i)})$ 、分散 σ^2 の正規分布に従う確率変数の実現値であるとする

- $y^{(i)}$ は $h(x^{(i)})$ に平均0、分散 σ^2 の正規分布に従うようなノイズ $\epsilon^{(i)}$ が合わさり観測されるとも考えられる
 - $y^{(i)} = h(x^{(i)}) + \epsilon^{(i)}$
- $\epsilon^{(i)}$ は独立同分布と仮定する

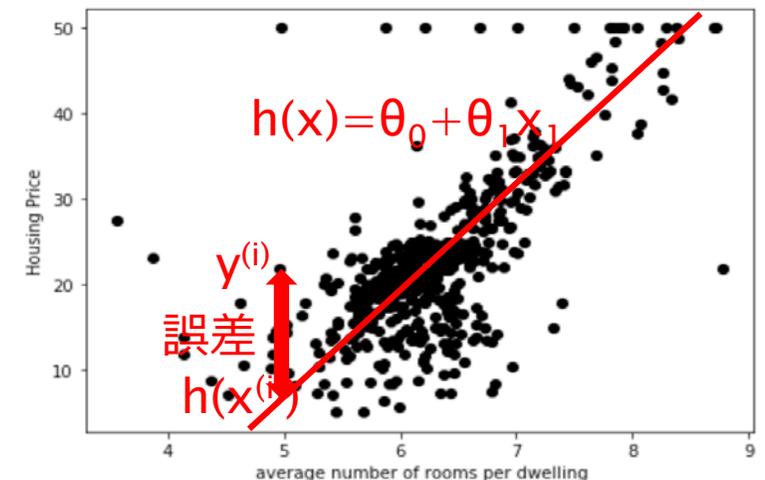
$\epsilon^{(i)}$ の確率密度関数（データ $x^{(i)}$ が与えられた時の $y^{(i)}$ の分布（パラメータは θ ））は

$$f(\epsilon^{(i)}) = f(y^{(i)} | x^{(i)}; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - h(x^{(i)}))^2}{2\sigma^2}\right)$$

最尤推定によりパラメータをする

対数尤度は

$$\log L(\theta) = \log \prod_{i=1}^m f(y^{(i)} | x^{(i)}; \theta) = \log \prod_{i=1}^m \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - h(x^{(i)}))^2}{2\sigma^2}\right)$$



復習：最尤推定による解釈

対数尤度（つづき）

$$\log L(\theta) = \log \prod_{i=1}^m \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - h(x^{(i)}))^2}{2\sigma^2}\right)$$

$$= \sum_{i=1}^m \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - h(x^{(i)}))^2}{2\sigma^2}\right)$$

$$= m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^m (y^{(i)} - h(x^{(i)}))^2$$

第1項は定数、対数尤度を最大にするには第2項を最小化すればよい

コスト関数に用いた残差平方和を最小化することに等しい $\sum_{i=1}^m (h(x^{(i)}) - y^{(i)})^2$

最小二乗法による線形回帰はパラメータの最尤推定をしているとも考えられる

正則化 MAP推定による解釈

訓練データセット $S = \{(x^{(i)}, y^{(i)})\}_{i=1}^m$ が与えられ時のパラメータの確率分布（事後分布）を考える

- ・ 訓練データセットを観測した後でどのようなパラメータが尤もらしいか

$$p(\theta|S) = \frac{p(S|\theta)p(\theta)}{p(S)}$$

- ・ $p(\theta)$ は訓練データセットを観測する前のパラメータの事前分布
 - ・ パラメータのあたりをつける事前知識

事後分布 $p(\theta|S)$ を最大化するようなパラメータが最適であるという事後確率最大化（MAP: Maximum a posteriori）に従ってパラメータを決めることを考える

$$\log p(\theta|S) = \log p(S|\theta) + \log p(\theta) - \log p(S)$$

対数尤度 $\log(p(S|\theta))$ と事前分布の対数 $\log(p(\theta))$ の和を最大にするようなパラメータを求めたい

MAP推定による正則化解釈

各特徴量のパラメータがそれぞれ平均0, 分散 σ^2 の正規分布に従うとする

- ・パラメータはなるべく小さい方がよいという事前知識

この時、事前分布の対数は

$$\log p(\theta) = -n \log \sqrt{2\pi\sigma} - \frac{1}{2\sigma^2} \|\theta\|_2^2$$

最尤推定による線形回帰の解釈（先週）より対数尤度の最大化はコスト関数の最小化なので

対数尤度 $\log(p(S|\theta))$ と事前分布の対数 $\log(p(\theta))$ の和を最大にするようなパラメータは

以下を最小化するようなパラメータ

$$\frac{1}{2\sigma^2} \sum_{i=1}^m (y^{(i)} - h(x^{(i)}))^2 + \frac{1}{2\sigma^2} \|\theta\|_2^2$$

コスト関数と正則化項の形になっている

その他の正則化

リッジ回帰

- 前述までのようなL2ノルムを正則化（L2正則化）に用いた線形回帰

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h(x^{(i)}) - y^{(i)})^2 + \lambda \|\theta\|_2^2$$

ラッソ回帰

- L1ノルムを正則化（L1正則化）に用いた線形回帰
- 解析解を持たない
 - 近接勾配法、座標降下法、交換方向乗数法などで推定する
- コスト関数を最小化する多くのパラメータが0となるため疎なパラメータ（重み）をえられる

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h(x^{(i)}) - y^{(i)})^2 + \lambda \|\theta\|_1$$

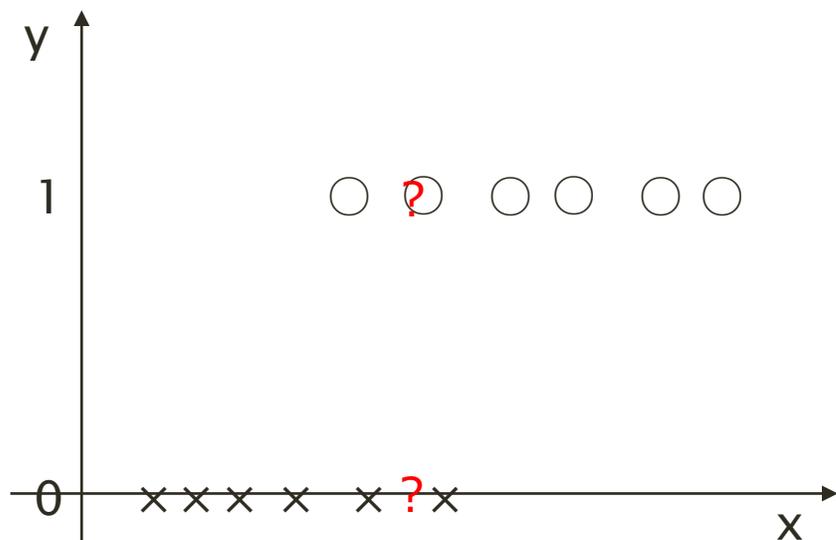
$$\|\theta\|_1 = |\theta_0| + |\theta_1| + \dots + |\theta_n|$$

判別問題

教師あり学習において出力が離散値

2値分類

- 出力が1か0
 - 出力の1か0は訓練データに対するラベルとも呼ばれる
 - 出力1のデータは正例、出力0のデータは負例とも呼ばれる



1変数入力の2値分類

ロジスティック回帰 シグモイド関数

1変数の入力に対して1か0の出力を予測する2値分類を考える

出力が1から0の範囲であるシグモイド関数 $g(z)$ によるフィッティングを考えてみる

- シグモイド関数は $z \rightarrow \infty$ の時は1, $z \rightarrow -\infty$ の時は0に近づく

$$g(z) = \frac{1}{1+e^{-z}}$$

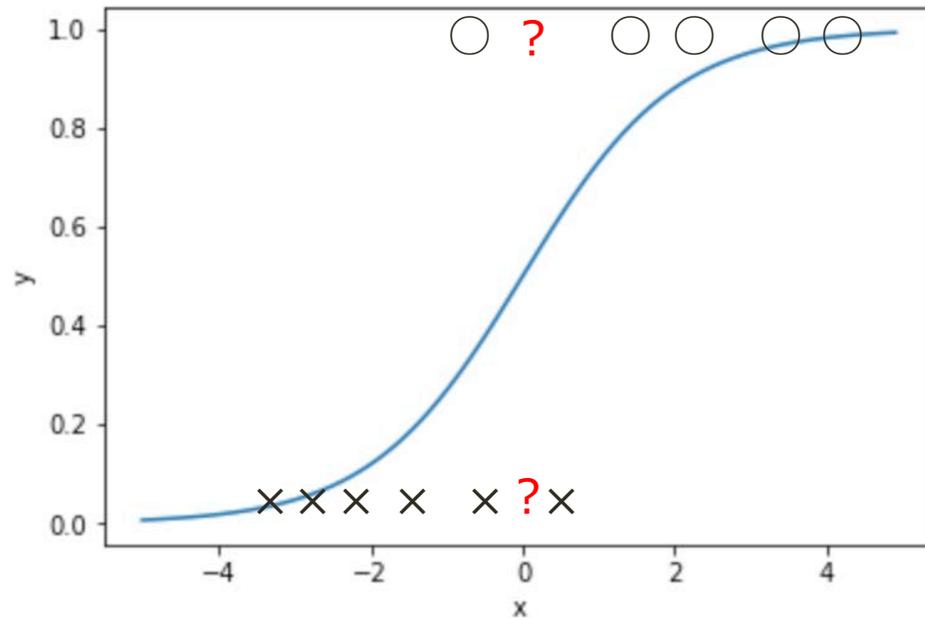
この時、入力 x とパラメータ θ について

仮説関数は以下のようになる

$$\theta = (\theta_0, \theta_1)^T$$

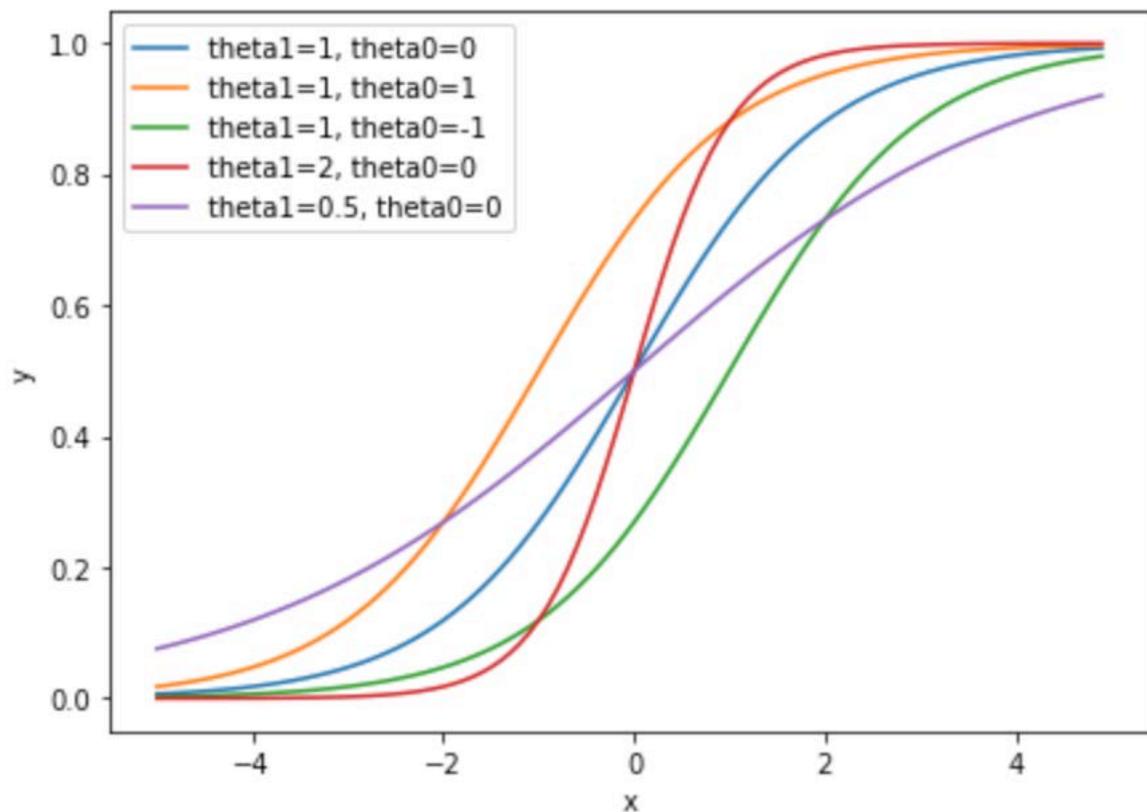
$$x = (x_0, x_1)^T$$

$$h(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$$

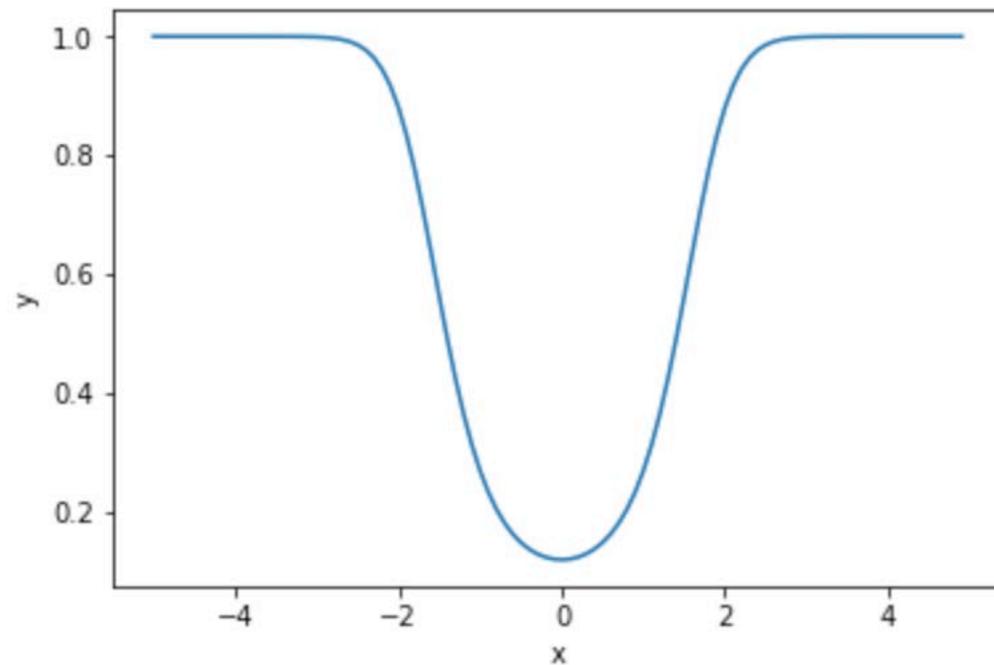


ロジスティック回帰 シグモイド関数

$$h(x) = g(\theta_0 + \theta_1 x) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x)}}$$



$$h(x) = g(\theta_0 + \theta_1 x^2) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x^2)}}$$



ロジスティック回帰

決定境界

シグモイド関数の出力 $h(x)=g(\theta_0+\theta_1x)$ はパラメータ θ の元で入力 x の出力 y が1であるかの確率の推定値 $P(y=1|x;\theta)$, $P(y=0|x;\theta)=1-P(y=1|x;\theta)$ と考えられる

訓練データセットからパラメータ θ^* が学習できれば以下のように分類ができる

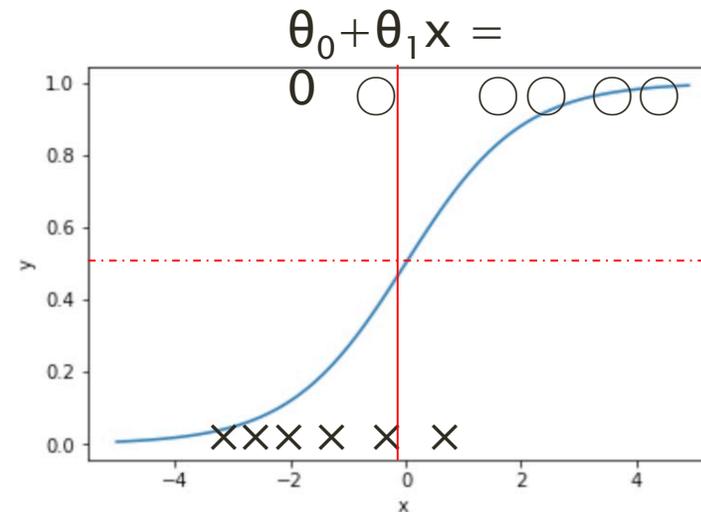
- $g(\theta_0+\theta_1x)$ が0.5以上 ($\theta_0+\theta_1x \geq 0$) であれば入力 x に対応する出力 (ラベル) は1
- $g(\theta_0+\theta_1x)$ が0.5未満 ($\theta_0+\theta_1x < 0$) であれば入力 x に対応する出力 (ラベル) は0

この時 $\theta_0+\theta_1x=0$ を決定境界と呼ぶ

一般に $g(\theta_0+\theta_1x)$ の閾値を c とすれば

$$c = g(\theta_0+\theta_1x) = 1 / (1 + \exp(-(\theta_0+\theta_1x)))$$

を満たす $\theta_0+\theta_1x$ が決定境界になる



ロジスティック回帰

コスト関数

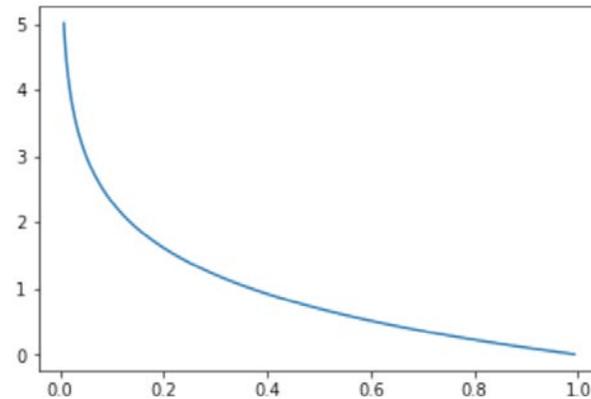
入力 x に対する仮説関数 $h(x)$ と出力（ラベル） y の誤差について以下のコスト関数を考えてみる

- y が1の時

仮説関数の出力（ y が1であるかの確率の推定値）
が小さいほど誤差は大きくなる

$$J(\theta_0, \theta_1) = -\log(g(\theta_0 + \theta_1 x))$$

$J(\theta_0, \theta_1)$



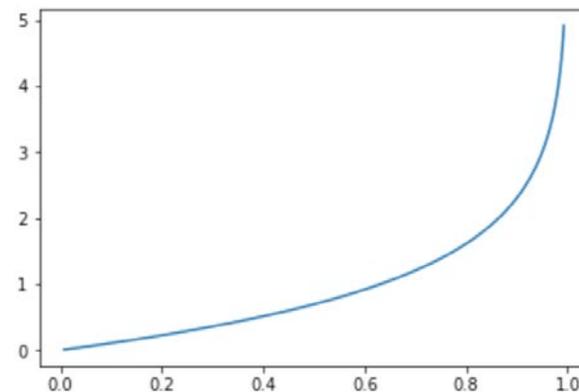
$g(\theta_0 + \theta_1 x)$

- y が0の時

仮説関数の出力（ y が1であるかの確率の推定値）
が大きいかほど誤差は大きくなる

$$J(\theta_0, \theta_1) = -\log(1 - g(\theta_0 + \theta_1 x))$$

$J(\theta_0, \theta_1)$



$g(\theta_0 + \theta_1 x)$

ロジスティック回帰 コスト関数の最小化

コスト関数

- 出力（ラベル） y が1の時、0の時を合わせて考えるとデータセット $\{(x^{(i)}, y^{(i)}); i=1, \dots, m\}$ についてコスト関数は

$$J(\theta_0, \theta_1) = -\frac{1}{m} \sum_{i=1}^m (y^{(i)} \log(g(\theta^T x^{(i)})) + (1 - y^{(i)}) \log(1 - g(\theta^T x^{(i)})))$$

- コスト関数は閉形式ではなく線形回帰のようにパラメータを解析的に求められない
- 最急降下法により訓練データセットを用いてコスト関数を最小化するようなパラメータを推定する

$$\theta_0 := \theta_0 - \alpha \frac{\partial J(\theta_0, \theta_1)}{\partial \theta_0} = \theta_0 - \alpha \sum_{i=1}^m (g(\theta^T x^{(i)}) - y^{(i)})/m$$

$$\theta_1 := \theta_1 - \alpha \frac{\partial J(\theta_0, \theta_1)}{\partial \theta_1} = \theta_1 - \alpha \sum_{i=1}^m ((g(\theta^T x^{(i)}) - y^{(i)})x^{(i)})/m$$

パラメータの更新式は線形回帰の時と同じ形になっている

* なぜ同じになるかは一般化線形モデル(GLM)で説明できる

補：コスト関数の偏微分

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{-1}{m} \sum_{i=1}^m (y^{(i)} \frac{\partial \log(h(x^{(i)}))}{\partial \theta_j} + (1 - y^{(i)}) \frac{\partial \log(1-h(x^{(i)}))}{\partial \theta_j})$$

$$= \frac{-1}{m} \sum_{i=1}^m (y^{(i)} \frac{\partial g(\theta^T x^{(i)})}{\partial \theta_j} / h(x^{(i)}) + (1 - y^{(i)}) \frac{\partial (1-g(\theta^T x^{(i)}))}{\partial \theta_j} / (1 - h(x^{(i)})))$$

$$= \frac{-1}{m} \sum_{i=1}^m (y^{(i)} (1 - h(x^{(i)})) x_j^{(i)} + (1 - y^{(i)}) h(x^{(i)}) x_j^{(i)})$$

$$= \frac{1}{m} \sum_{i=1}^m (h(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

この式展開で以下を用いた

$$g'(x) = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}} \left(1 - \frac{1}{1+e^{-x}}\right) = g(x)(1 - g(x))$$

ロジスティック回帰 パラメータの推定

入力の変数が多変数（ n 次元の特徴量ベクトル）の時も同様に以下の様に各パラメータ θ_j を更新してコスト関数 $J(\theta)$ が最小となるパラメータを求めればよい

$$\theta_j := \theta_j - \alpha \frac{\partial J(\theta)}{\partial \theta_j} = \theta_j - \alpha \sum_{i=1}^m ((g(\theta^T x^{(i)}) - y^{(i)}) x_j^{(i)}) / m$$

データセットの入力 X と正解 y を元にコスト関数 $J(\theta)$ を最小にするようなパラメータ θ を求める更新は以下のように表すことができる

$$\theta := \theta - \frac{\alpha}{m} X^T (g(X\theta) - y) \quad X = \begin{pmatrix} x_0^{(1)} & x_1^{(1)} & \dots & x_n^{(1)} \\ \dots & \dots & \dots & \dots \\ x_0^{(m)} & x_1^{(m)} & \dots & x_n^{(m)} \end{pmatrix} \quad y = (y^{(1)}, y^{(2)}, \dots, y^{(m)})^T$$
$$\theta = (\theta_0, \theta_1, \dots, \theta_n)^T$$

ロジスティック回帰 最尤推定による解釈

シグモイド関数の出力 $h(x)=g(\theta^T x)$ はパラメータ θ の元で入力 x の出力 y が1であるかの確率の推定値であった

- $P(y=1|x;\theta)=h(x)$
- $P(y=0|x;\theta)=1-h(x)$

この確率の分布は以下のように表せる（ベルヌーイ分布）

$$p(y|x; \theta) = (h(x))^y(1 - h(x))^{1-y}$$

最尤推定によりパラメータを推定する。対数尤度は

$$\log L(\theta) = \prod_{i=1}^m (h(x^{(i)}))^{y^{(i)}}(1 - h(x^{(i)}))^{1-y^{(i)}}$$

$$= \sum_{i=1}^m (\log y^{(i)} h(x^{(i)}) + (1 - y^{(i)}) \log(1 - h(x^{(i)}))) \quad \text{ロジスティック回帰のコスト関数となっている}$$

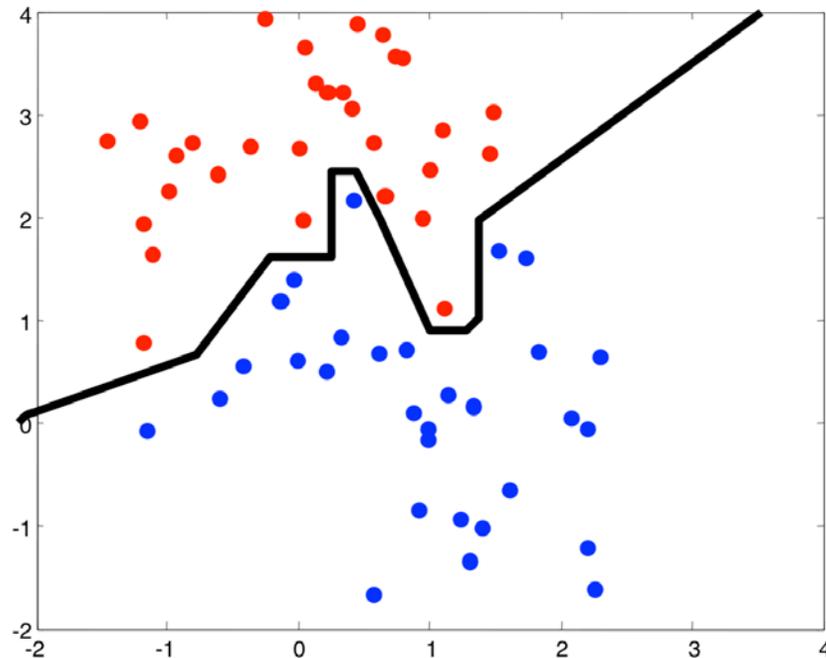
ロジスティック回帰

正則化

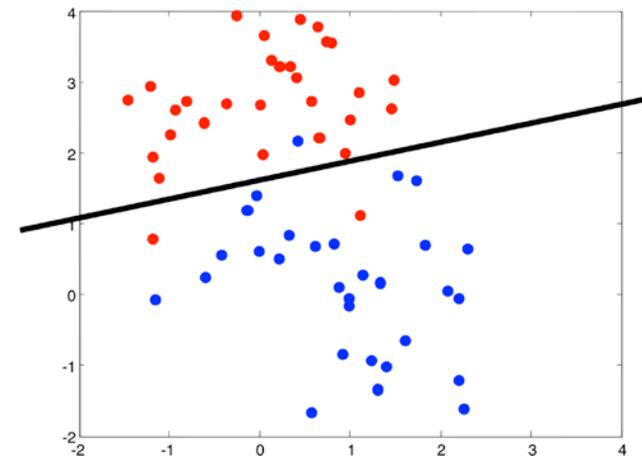
線形回帰と同様にロジスティック回帰でも学習モデルが複雑なほど過学習が起こりやすい

- 学習モデルが複雑であればより複雑な決定境界がえられる

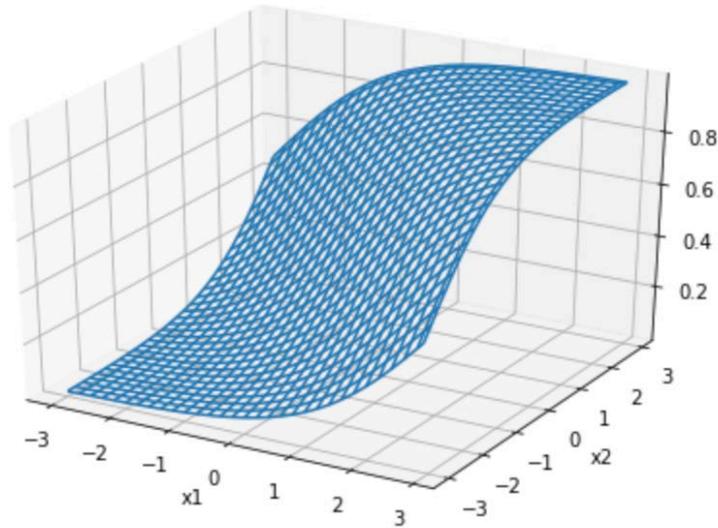
2変数入力, 2クラス分類の決定境界 (複雑な学習モデル)



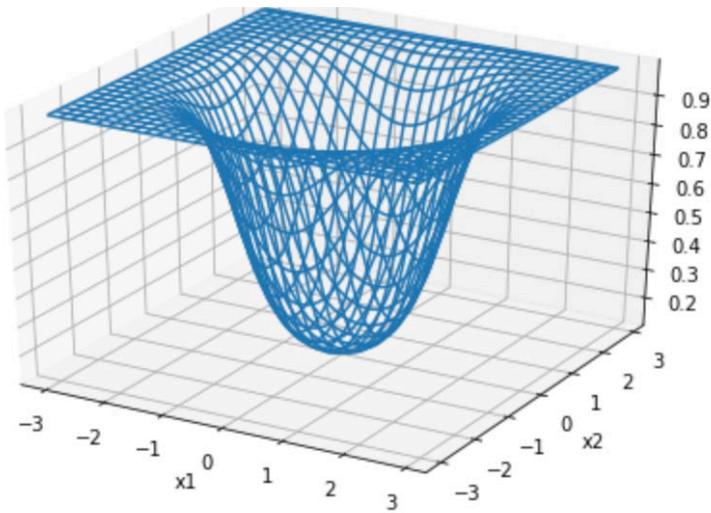
2変数入力, 2クラス分類の決定境界 (単純なモデル)



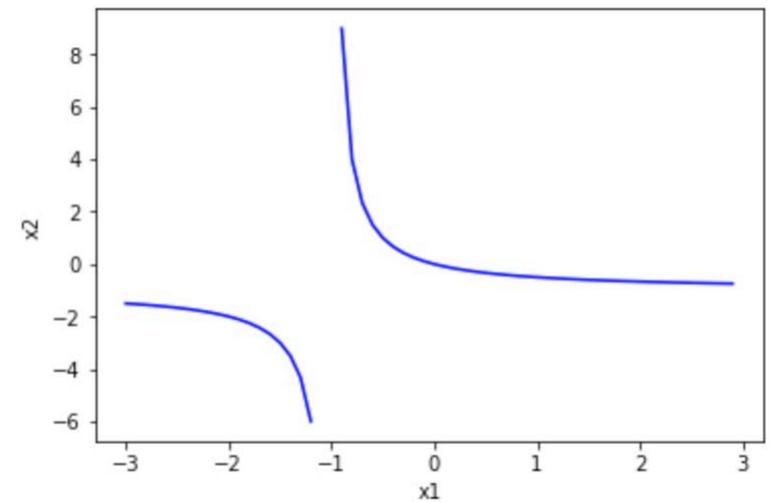
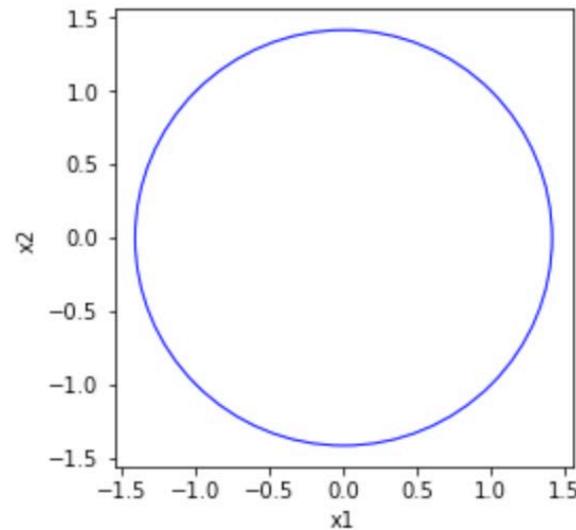
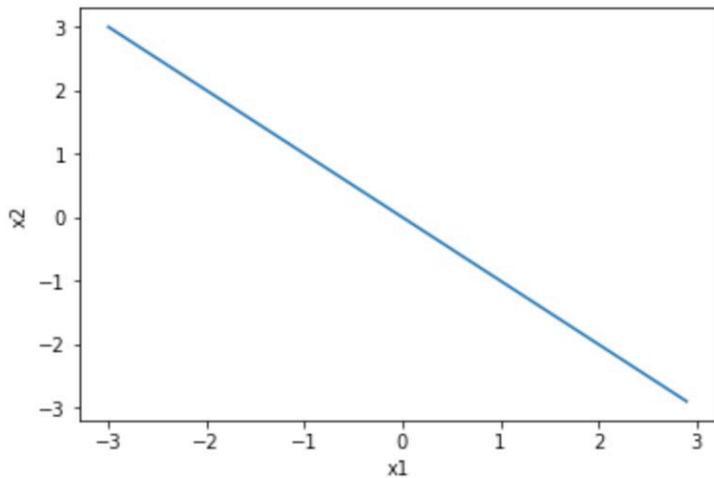
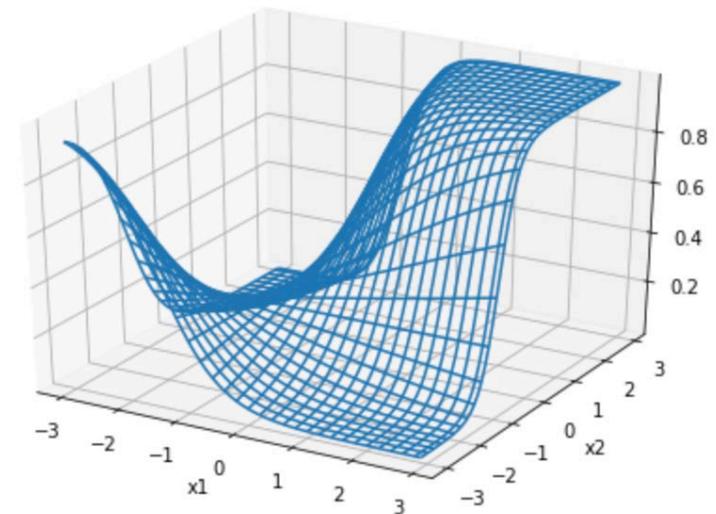
$$h(x) = \frac{1}{1+e^{-(\theta_0+\theta_1x_1+\theta_2x_2)}}$$



$$h(x) = \frac{1}{1+e^{-(\theta_0+\theta_1x_1^2+\theta_2x_2^2)}}$$



$$h(x) = \frac{1}{1+e^{-(\theta_0+\theta_1x_1+\theta_2x_2+\theta_3x_1x_2)}}$$



ロジスティック回帰

正則化

線形回帰と同様に、過学習（パラメータの肥大化）が起こったらペナルティを課すような項をコスト関数に追加する

- ・ λ はハイパーパラメータ（大きいほど過学習を抑える）

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^m (y^{(i)} \log(g(\theta^T x^{(i)})) + (1 - y^{(i)}) \log(1 - g(\theta^T x^{(i)}))) + \frac{\lambda}{2m} \theta^T \theta$$

* θ_0 は含めない

正則化を用いた最急勾配法によるパラメータ更新

- ・ バイアス項に対するパラメータ θ_0 へはペナルティを与えないようにする

$$\theta_j := \theta_j - \alpha \frac{\partial J(\theta)}{\partial \theta_j} = \theta_j (1 - \alpha \frac{\lambda}{m}) - \alpha \sum_{i=1}^m ((h(x^{(i)}) - y^{(i)}) x_j^{(i)}) / m$$

$$\theta_0 := \theta_0 - \alpha \sum_{i=1}^m (h(x^{(i)}) - y^{(i)}) / m$$

ロジスティック回帰 多クラス分類

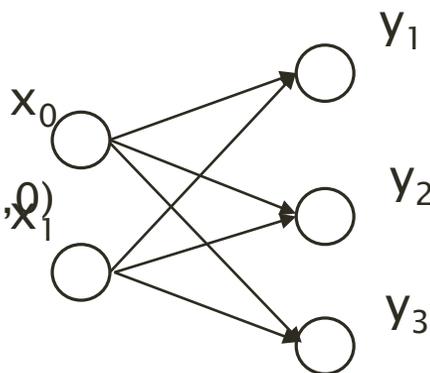
出力 y が多クラスの場合、入力 x に対して y があるクラス c である確率の推定値を以下のように考える

- ソフトマックス関数

- 例：出力 y が3クラスの場合

- $y = (y_1, y_2, y_3) = (1, 0, 0)$ or $(0, 0, 1)$ or $(0, 1, 0)$

$$P(Y = y_c | x) = \frac{\exp(\theta_c^T x)}{\sum_i \exp(\theta_i^T x)}$$



\exp が大きくなりすぎないように
実装上は以下のように $\log C$ で正規化

$$\frac{C \exp(\theta_c^T x)}{C \sum_i \exp(\theta_i^T x)} = \frac{\exp(\theta_c^T x + \log C)}{\sum_i \exp(\theta_i^T x + \log C)}$$

$$\log C = -\operatorname{argmax}_c \theta_c^T x$$

この時、以下のコスト関数を最小化するパラメータを推定する

- パラメータ数は(入力変数の数) \times (クラスの数)

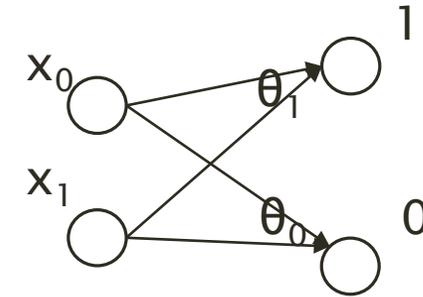
$$J(\theta) = -\frac{1}{m} \sum_{i=1}^m \sum_c y_c^{(i)} \log(p(y_c^{(i)} | x^{(i)}))$$

$$\frac{\partial J(\theta)}{\partial \theta_c} = \frac{1}{m} \sum_{i=1}^m (p(y_c^{(i)} | x^{(i)}) - y_c^{(i)}) x^{(i)}$$

ロジスティック回帰 多クラス分類

ソフトマックス関数で2クラス分類を考えてみる

$$P(y = 1|x) = \frac{\exp(\theta_1^T x)}{\exp(\theta_1^T x) + \exp(\theta_0^T x)}, P(y = 0|x) = \frac{\exp(\theta_0^T x)}{\exp(\theta_1^T x) + \exp(\theta_0^T x)}$$



$\exp(\theta_1^T x)$ で割ると

$$P(y = 1|x) = \frac{1}{1 + \exp((\theta_0 - \theta_1)^T x)}, P(y = 0|x) = \frac{\exp((\theta_0 - \theta_1)^T x)}{1 + \exp((\theta_0 - \theta_1)^T x)}$$

$\theta_0 - \theta_1$ を θ とおくとロジスティック回帰の仮説関数となる

$$P(y = 1|x) = \frac{1}{1 + \exp(\theta^T x)},$$

$$P(y = 0|x) = 1 - P(y = 1|x)$$

参考書

著作権の都合により
ここに挿入されていた画像を削除しました

書籍『Machine Learning :
A Probabilistic Perspective』表紙
Kevin P. Murphy 著
MIT Press (2012.8)
<https://mitpress.mit.edu/books/machine-learning-1>

著作権の都合により
ここに挿入されていた画像を削除しました

書籍『統計的学習の基礎—データマイニング・推論・予測—』表紙
Trevor Hastie ・ Robert Tibshirani ・
Jerome Friedman 著
杉山 将 ・ 井手 剛 ・ 神尾 敏弘 ・ 栗田 多喜夫 ・
前田 英作 監訳
井尻 善久 ・ 井手 剛 ・ 岩田 具治 ・ 金森 敬文 ・
兼村 厚範 ・ 烏山 昌幸 ・ 河原 吉伸 ・ 木村 昭
悟 ・ 小西 嘉典 ・ 酒井 智弥 ・ 鈴木 大慈 ・ 竹内
一郎 ・ 玉木 徹 ・ 出口 大輔 ・ 富岡 亮太 ・ 波部
斉 ・ 前田 新一 ・ 持橋 大地 ・ 山田 誠 訳
共立出版 (2014.6)
<https://www.kyoritsu-pub.co.jp/bookdetail/9784320123625>

著作権の都合により
ここに挿入されていた画像を削除しました

書籍『パターン認識と機械学習 上
ベイズ理論による統計的予測』表紙
Christopher M. Bishop 著
栗田 多喜夫 ・ 樋口 知之 ・
松本 裕治 ・ 村田 昇 監訳
丸善出版(2012.1)
<https://www.maruzen-publishing.co.jp/item/b294524.html>