PRML Seminar #1

1.1-1.6.1 #PRML 学ぼう

Shunya Ueta

Graduate School of SIE, Univ. of Tsukuba
Department of Computer Science

April 10, 2015

自己紹介

- **名前**:**上田隼也** (@hurutoriya)
- **筑波大学大学院** 1 **年** Go to Doctor course :)
- **情報数理研究室所属**
- **研究分野** : **画像認識・機械学習**

この勉強会について

- パターン認識と機械学習についての輪講です
  機械学習とパターン認識の基礎を理解、実用レベルで使いこなす事を目的にセミナーを開催していきます
- 2015 年を目処に一周予定
- 受講者には基礎的な微積分・線形代数・確率統計の知識を前提としています
- 資料中のサンプルコードは Python を採用しています。
- 勉強会に関する情報については Hashtag: #PRML 学ぼう を使って発信していきます

今回の担当

$1 \rightarrow 1.6$

機械学習とは?

パターン認識 (Pattern Recognition):
計算機アルゴリズムを通じて、データの中の規則を自動的に見つけ出す。更にその規則性を用いてデータを異なるカテゴリに分類する。
例) 手書き数字の認識
入力として 28×28 の大きさの手書き数字の画像がある。入力データは 784 次元の実数値ベクトル x で表現できる。ベクトル x を入力として受け取り、それが $0 \ldots 9$ のどの数字を表しているかを出力する機械を作る。

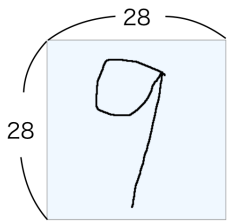


図 1:　手書き数字の画像例

実現方法 (1)

人力によるルールの作成 $\rightarrow$ ルール数の発散 (実現不可)

機械学習的アプローチ

- **訓練集合** (training set):
  N **個の手書き文字の大きな集合を用意する** $\{x_1, \ldots, x_N\}$
- **目標ベクトル** (target vector): t
  一つ一つの数字に対応するカテゴリを表すベクトル

最終的に得られるのは $y(x)$ **である。**
この関数に数字の画像 x **を入力すると、目標ベクトルと画像データに符号化可能なデータが合成されたベクトル** y**(目標ベクトル)がラベリングされた画像が返ってくる。**

実現方法 (2)

- **訓練** (training), **学習** (leaning) **段階** :
  training set **のみでモデル化されている状態**
- **テスト集合** (test set) :
  訓練集合以外のデータ (**未知のデータ**)
- **汎化** (generalization) :
  訓練集合以外のデータ (**未知のデータ**) **に対して適応可能にさせること**

実問題として入力データは大きな多様性を持つ。$\rightarrow$
汎化が中心的な課題となる。

例)Ajax を使った手書き文字認識

図 2:　Ajax を使った手書き文字認識

参考: http://chasen.org/ taku/software/ajax/hwr/

前処理 (Preprocessing)

実世界では、入力変数は**前処理** (Preprocessing) により問題を解きやすくしておく。

例) 手書き数字

- 数字画像に変形 (アフィン変換)、拡大・縮小を行い
  同一の大きさに変換 $\rightarrow$ 入力データの多様性の減少

前処理の段階は特徴抽出 (feature extraction) と呼ばれる。
多様性減少の目的以外にも、計算の高速化のためにも用いられる事が多い。

機械学習の分類

1. **教師あり学習** (supervised learning) : **訓練データがラベリングされている状態での問題**
   - **クラス分類** (classification) **問題** : **各入力ベクトルを有限個の離散カテゴリに分類する問題**
   - **回帰** (regression) : **求める出力が一つないしはそれ以上の連続変数であるような問題**
2. **教師なし学習** (unsupervised learning) : **訓練データがラベリングされている状態での問題**
   - **クラスタリング** (clustering) : **類似した事例のグループを見つける**
   - **密度推定** (density estimation) : **入力空間におけるデータの分布を見つける**
3. **半教師あり学習** (semis-upervised learning) : **訓練データがラベリングされているものと非ラベリング状態の物が混在している状態での問題**

強化学習 (reinforcement learning)(1)

教科学習 (reinforcement learning) : **ある与えられた条件下で、報酬を最大化するような適当な行動を見つける問題。**
状態と行動の系列から環境との相互作用を通じて学習を行う (行動基準は直近の報酬だけではなく、過去の行動も参考にされる)。
教師あり学習との違い : 最適な答えは与えられずに試行錯誤を通じて学習アルゴリズム自らが最適解を発見する

バックギャモンに対する強化学習の適用 (Tesauro 1994)

ニューラルネットワーク (第 5 章) により、自分自身のコピーと何百万ものゲームをこなす必要がある。
選択肢は無数に存在するが、勝利という形でしか報酬を与えることができない。そのため、勝利に関係する手に対しては正確に報酬を割り当てる必要がある (信頼度割り当て問題)。

強化学習 (reinforcement learning)(2)

下記の２つを行い強化学習を行う (トレードオフの関係)。

- **探査** (exploration) : **新規の手がどれほど有効なのかを探す**
- **利用** (exploitation) : **高い報酬が得られることがわかっている行動を取る**

第 1 章で導入する 3 つの重要な道具

1. **確率論**
2. **決定理論**
3. **情報理論**

1.1　多項式曲線フィッティング

訓練データ
入力： N **個の観測値** x **を並べた** $\boldsymbol{x} = (x_1, \ldots, x_N)^T$
出力：それぞれに対応する観測値 $\boldsymbol{t} = (t_1, \ldots, t_N)^T$
未知の入力変数 x **に対して目標変数を予測したい** t **(汎化)**
目標とするモデルは $\sin 2\pi x$**、訓練データはノイズを乗せて**
$N = 10$ **で与えられる。**

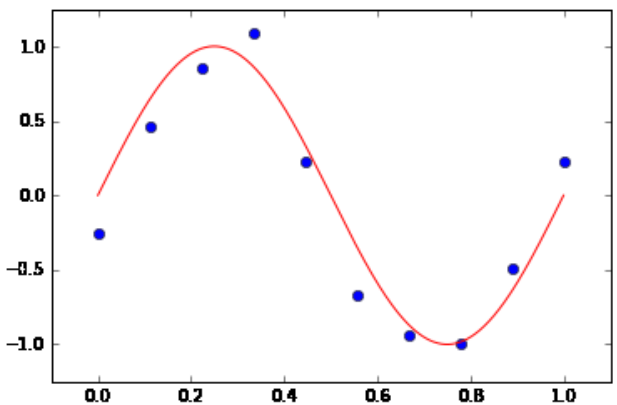


図 3:　ex)N=10 **で訓練データが与えられている**

1.1 多項式曲線フィッティング

$$y(x, \boldsymbol{w}) = w_0 x^0 + w_1 x^1 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

M:**多項式の次数** (order)
ω:**多項式の係数、まとめて** ω **で表す**

$$E(\boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \boldsymbol{w}) - t_n\}^2$$

訓練データに多項式をあてはめることで、係数の値を求めたい $\rightarrow$
誤差関数 (Error Function) **の最小化を目指す。**

Fitting Order Case:M=0,1

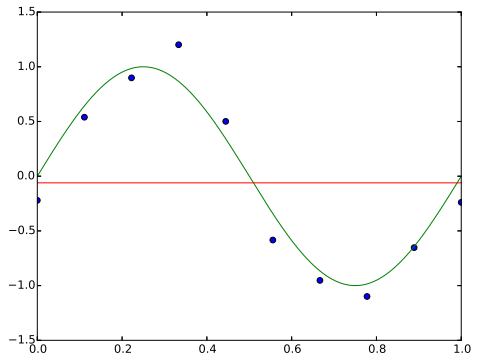


図 4:　M=0

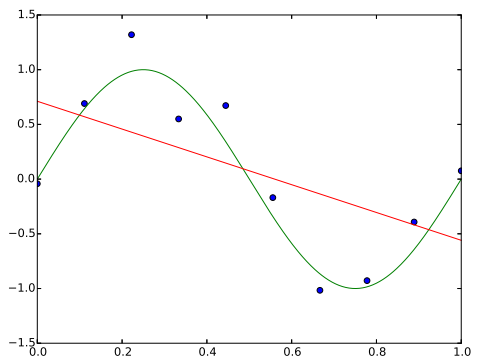


図 5:　M=1

Fitting Order Case:M=3,9

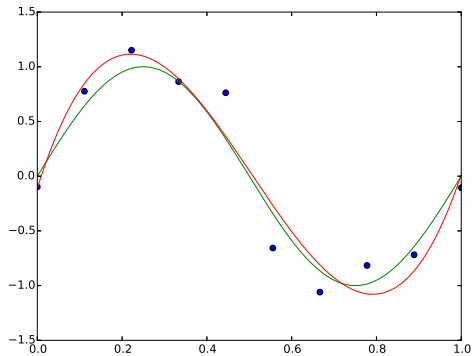


図 6:　M=3

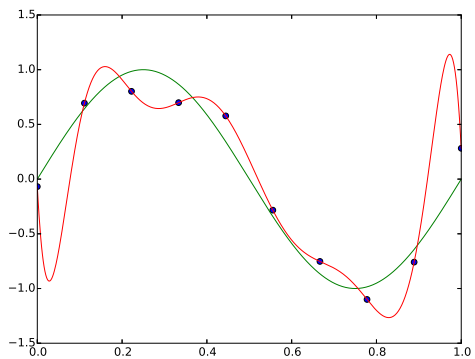


図 7:　M=9

過学習と平均二乗平方根誤差

9 次でフィッティングした場合、**過学習** (over-fitting)
機械学習の目標:未知のデータに対して精度の高い予測 (汎化)

$$E_{RMS} = \sqrt{2E(\boldsymbol{w})/N}$$

平均二乗平方根誤差 (root-mean-square error,RMS error)

- N で割ることでサンプル数のギャップを消す
- 平方根を取ることで、元の尺度に戻す

データセットのサイズに応じた過学習の様子

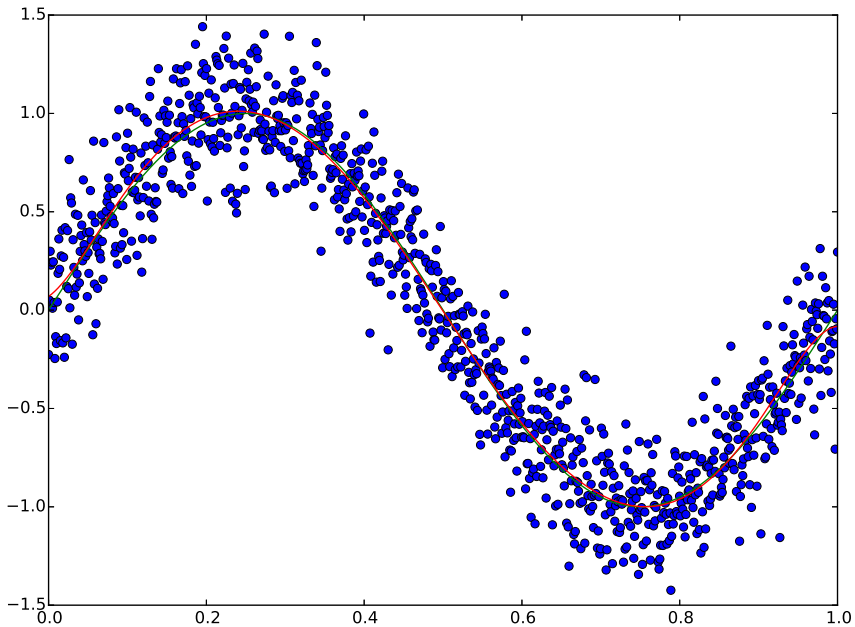


図 8:　M=9,N=1000

誤差関数の正則化 (regularization) 過学習を制御

$$E(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{\omega}) - t_n\}^2 + \frac{\lambda}{2}||\boldsymbol{w}||^2$$

２次の正則化の場合、リッジ回帰 (ridge regression)

検証用集合 (Validation set):$\boldsymbol{w}$ **を決定するためにデータセット**
ホールド・アウト集合 (hold-out set) **とも呼ばれる。**
欠点: **貴重なデータを無駄にする**

確率論

データにはノイズが必ず付随し、データセットのサイズも有限である。**不確実性**が重要な概念となる。

確率論の概念を簡単に説明

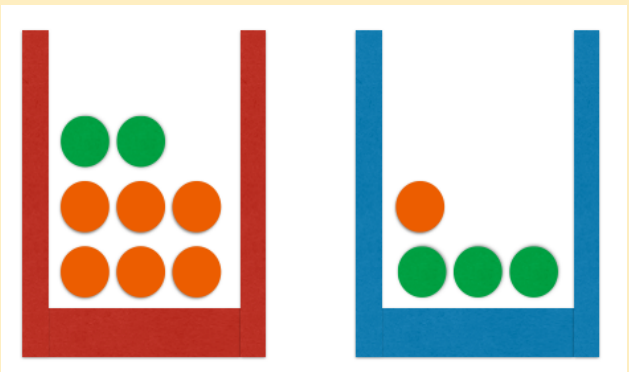


図 9:　赤と青、２つの箱がある。赤の箱にはりんごが 2 個、オレンジが 6 個、青の箱にはりんごが 3 個、オレンジが一個入っている。赤の箱を 40%, 青の箱を 60%で選び、果物は同じ確からしさで選ぶ。

確率の基本的法則

確率の加法定理 (Sum rule)

$$p(X) = \sum_{Y} p(X, Y)$$

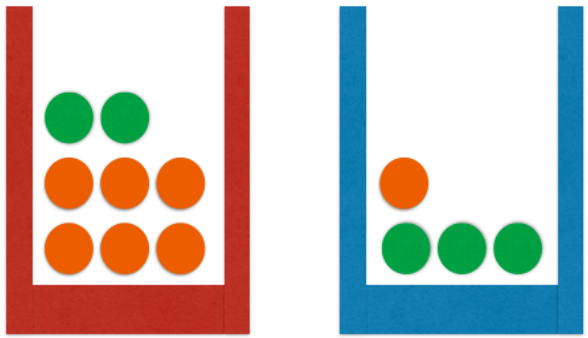
確率の乗法定理 (Prduct rule)

$$p(X, Y) = p(Y|X)p(X)$$

ベイズの定理 (Bay's theorem)

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

ベイズの定理の直感的な説明



どの箱を選びましたか?

事前確率 (prior probability)

事前に得られる確率値 $p(Box)$

事後確率 (posterior probability)

果物を選んだ後確定する確率 $p(Box|Fruit)$

一旦果物がオレンジだとわかれば、赤い箱はオレンジの数が多い
→ 赤い箱である確率が高くなる。

1.2.1 連続変数への拡張

確率密度 (probability density) **対象:連続値** (continuous value)
連続値である変数 x **が区間** $(x, x + \delta x)$ **に入る確率が** $\delta x \to 0$ **で与えられた時の** x **上の** $p(x)$

1. $p(x) \geq 0$
2. $\int_{-\infty}^{\infty} p(x)dx = 1$

確率質量 (probability mass) **対象:離散集合** (discrete set)
離散変数である x **がある** x **になる確率** $p(x)$

1.2.2 期待値と分散について

期待値 (expectation)

ある関数 $f(x)$ の確率分布 $p(x)$ 下での平均値

離散分布 $\mathbb{E}[f] = \sum_x p(x)f(x)$

連続変数 $\mathbb{E}[f] = \int p(x)f(x)dx$

分散 (variance)

$$
\begin{aligned}
\mathrm{var[f]} &= \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] \\
&= \mathbb{E}[f(x)^2 - 2f(x)(E)[f(x)] + \mathbb{E}[f(x)]^2] \\
&= \mathbb{E}[f(x)^2 - 2f(x)(E)[f(x)] + \mathbb{E}[f(x)]^2] \\
&= \mathbb{E}[f(x)^2] - 2\mathbb{E}[f(x)](E)[f(x)] + \mathbb{E}[f(x)]^2 \\
&= \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2 \quad \text{seminar1.5}
\end{aligned}
$$

共分散 (covariance)

$$
\begin{aligned}
\mathrm{cov}[\mathrm{x}, \mathrm{y}] &= \mathbb{E}_{x,y}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] \\
&= \mathbb{E}_{x,y}[xy + \mathbb{E}[x]\mathbb{E}[y] - x\mathbb{E}[y] - y\mathbb{E}[x]] \\
&= \mathbb{E}_{x,y}[xy] + \mathbb{E}_{x,y}[\mathbb{E}[x]\mathbb{E}[y]] - \mathbb{E}_{x,y}[x\mathbb{E}[y]] - \mathbb{E}_{x,y}[y\mathbb{E}[x]] \\
&= \mathbb{E}_{x,y}[xy] + \mathbb{E}[x]\mathbb{E}[y] - \mathbb{E}[x]\mathbb{E}[y]] - \mathbb{E}[y]\mathbb{E}[x] \\
&= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] \\
&= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y](x, y\text{: independ}) \\
&= \mathbb{E}[x]\mathbb{E}[y] - \mathbb{E}[x]\mathbb{E}[y](x, y\text{: independ}) \\
&= 0 \quad \text{seminar1.6}
\end{aligned}
$$

1.2.3 Frequentist VS Bayesian

Frequentist

確率をランダムな繰り返し的試行の頻度とみなす

Bayesian

不確実性の度合いを確率とする
例) この世界に何度も繰り返し行うことができる事象がどれだけあるか考えてみて欲しい。
南極の氷が喪失するなど不確かな事象が起きたとしよう。年間どの程度溶けているかの情報を得ることで、その情報を適応することで南極の氷が喪失する不確かさを予測する。

1.2.3 Likelifood function working on Frequentist, Bayesian

尤度関数 (likelihood function)

$p(Data|param)$　**データに対する評価。パラメータの関数とみなせる。**

Frequentist

パラメータは固定されているものと考えられている。データの分布を考慮して、パラメータは決定される。

Bayesian

データは唯一に定まり、パラメータに関する不確実性は w の確率分布として表現される。

Bayesian の利点

事前知識を自然に導入できること

例) Frequentist における観測に偏りが発生した場合の確率

公平に表・裏がでるコインを 3 回投げて毎回表がでた。古典的な最尤推定では、表が出る確率は 1 になってしまう。

尤度

尤度は確率と数値的に同じである。
例) サイコロを振って 1 が三回連続同じ物が出る同時確率と尤度は同値である。
違いは、確率は「事象の確率」であり、尤度は「観測データ下での仮説の尤度」である。
(likelihood for a hypothesis given a set of observations)

1.2.4 Gaussian distribution

$$N(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

params: μ, σ^2

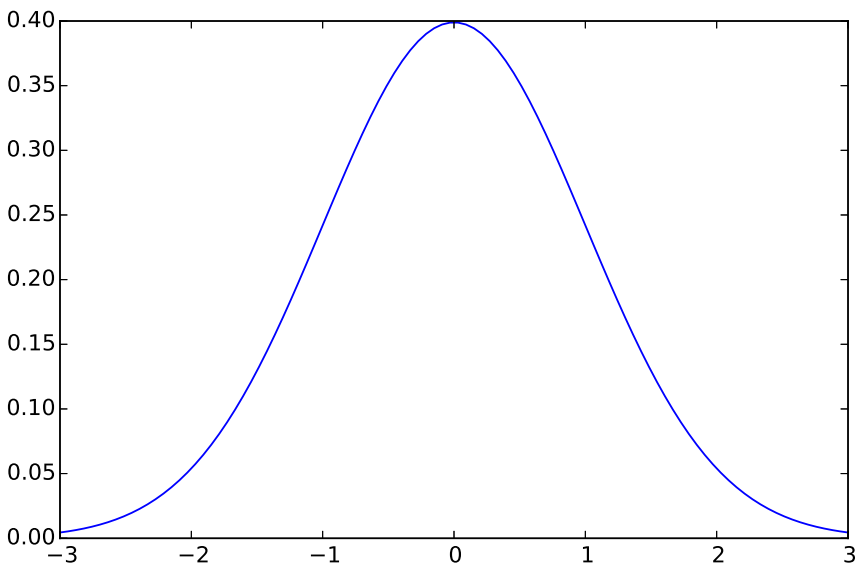


図 10:　$\mu = 0, \sigma = 1$

1.2.5 Re:Fitting curve

Bayesian **における曲線フィッティング**

$$p(t|x, \boldsymbol{w}, \beta) = N(t|y(x, \boldsymbol{w}), \beta^{-1})$$

分布の逆分散に相当するパラメータ β **を定義。**
訓練データ x, t **を使って、未知のパラメータ** w, β **を求めるのに**
最尤推定を用いる。

$$p(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} N(t_n|y(x_n, \boldsymbol{w}), \beta^{-1})$$

Maximum likelifood function

尤度関数を最大化するさいには、対数を用いる。利点として

1. **乗算が加算に変化**
2. **確率は少数で表現されるので、乗算を行うとアンダーフローが頻発**

$$p(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} N(t_n|y(x_n, \boldsymbol{w}), \beta^{-1})$$

$$\ln p(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{N}{2}, \ln \beta - \frac{N}{2} ln 2\pi$$

予測分布 (predictive distribution)

パラメータベクトル $\boldsymbol{w}_{ML}$ をまず決定し、β_{ML} を推定する。

$$\frac{1}{\beta_{ML}} = \frac{1}{n}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}_{ML}) - t_n\}^2$$

β_{ML} が決定されたことにより予測分布という形で t の確率分布を考えることができる。

$$p(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} N(t_n|y(x_n, \boldsymbol{w}), \beta^{-1})$$

実世界問題への応用

曲線フィッティングでは、入力パラメータは一つ。現実問題では入力パラメータは複数個存在するのが当たり前である。
自分たちは 3 次元の存在。4 次元以上の空間に関しては幾何的直感は働きづらい。

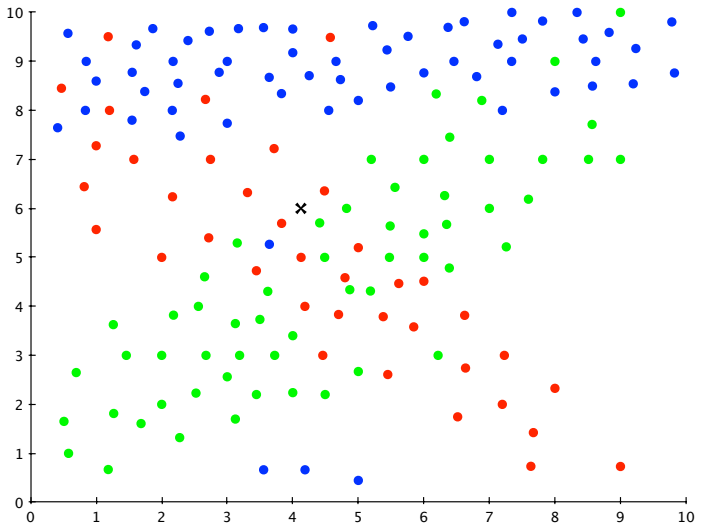


図 11:　3 クラス存在するデータセットの散布図、わかりやすいように２次元の部分空間へ

割り当て方法の提案

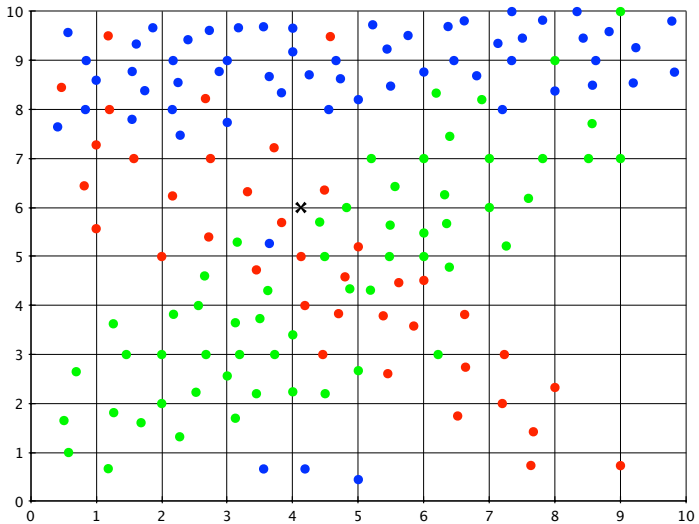


図 12:　同グリッド内で最も多いクラスを割り当てる

高次元に有効な２つの性質

1. **実データでも有効なデータは低次元に集中している**
2. **データは滑らかな事が多い (局所的に) ので、内挿により対応する**
   学習モデルを適用する必要はなく、内挿で事足りる

About decision theory

入力ベクトル x と目標変数 t があり、x の新たな値に対して対応する t を求めたい。

回帰問題 t は連続変数

クラス分類 t はクラスラベル

推論 (inference) 訓練データ x, t から同時確率分布 $p(x, t)$ を求めること

決定理論の主題は推論で得られた確率分布 t を予測し決定すること。

key of decision theory

- Berger, James O. Statistical decision theory and Bayesian analysis. Springer Science & Business Media, 1985.
- Bather, John. "Decision Theory.A n Introduction to Dynamic Programming and Sequential Decisions." (2000). APA

例) 患者の X 線画像をベクトル化した x を使用して、その患者が癌かどうかをクラス C_1(癌患者),C_2(非癌患者) にわけたい。
($p(C_k|\boldsymbol{x}$ を求めたい)

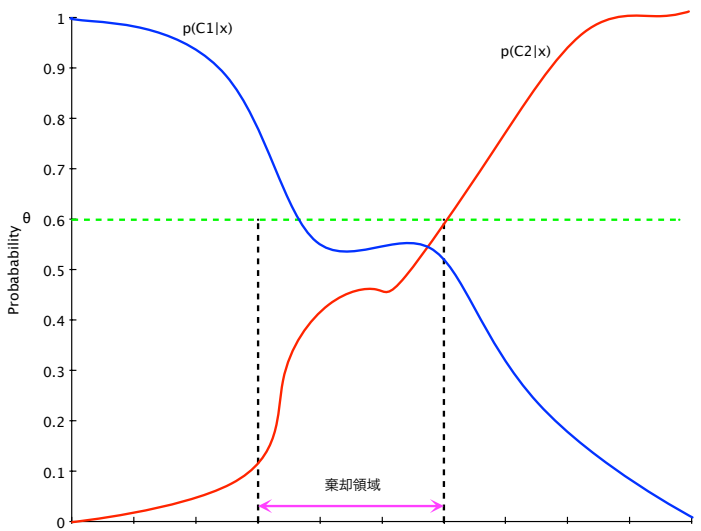
$$p(C_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_k)p(C_k)}{p(\boldsymbol{x})}$$

$p(C_k)$ は X 線画像を適用する前に人間が癌にかかる事前確率、
$p(C_k|\boldsymbol{x})$ は X 線画像から得られた情報を使用してベイズの定理を用いて修正した事後確率である。

some decision rule

損失関数 (loss function) **決定・行動に付随する損失を表す関数**
癌患者のクラス分類で、どちらが患者により大きい損失を与えるだろうか?
- **癌でない人を癌と診断する**
- **癌の人を癌と診断する**

棄却オプション (reject option) **決定が困難な場合には、決定を避ける選択**



1.5.4 inferense and decision

推論段階 → 決定段階への３つのアプローチ

生成モデル (generative model)
クラスごとにクラスの条件付き密度 $p(x|C_k)$ を決定する推論問題を解く。そして $p(C_k)$ も求めて、事後確率 $p(C_k|x)$ を求める。そして決定理論によりクラスに割り当てる。出力だけでなく入力もモデル化されるので、データの生成も可能になる。

識別モデル (discriminative model)
クラス事後確率 $p(C_k|x)$ を決定する推論問題を解き、決定理論を用いて新たな x をクラスに割り当てる。

識別関数 (discriminative function) **推論せずに直接決定する。各入力 x から直接クラスラベルへと写像する。**

情報量とエントロピー

ある離散確率変数 x があるとする。
$h(x) = -\log_2 p(x)$: **情報量を表す**
$H[x] = -\sum_x p(x) \log_2 p(x)$: **確率変数 x のエントロピー**

**確率変数 x を与え、8 個の可能な状態を等確率で取る。
その値を受信者に 3bit の長さにして送る。
この変数のエントロピーは、**

$$H[x] = -8 \times \frac{1}{8} \log_2 \frac{1}{8} = 3bit$$

**次に x は a, b, c, d, e, f, g の 8 この可能な状態をとり、
$\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\}$ の確率で与えられる。その際のエントロピーは**

$$\begin{aligned} H[x] &= -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{8} \log_2 \frac{1}{8} - \frac{1}{16} \log_2 \frac{1}{16} - \frac{4}{64} \log_2 \frac{1}{64} \\ &= 2bit \end{aligned}$$

確率の分布が非一様なものよりも、一様な分布のほうがエントロピーは高い事がわかる。エントロピーは確率分布と密接に関係がある。