クラシックな機械学習の入門

4.　学習データと予測性能 Bias^2 - Variance - Noise 分解

過学習

損失関数とBias,Variance, Noise

K-Nearest Neighbor法への応用

bias2とvarianceの間のトレードオフの

線形回帰への応用

by 中川裕志（東京大学）

過学習：over-fitting

- 教師データによる学習の目的は未知のデータの正確な分類や識別
- 過学習(over-fitting)
  - 教師データに追従しようとすればするほど、複雑なモデル（＝パラメタ数の多い）になり、教師データへの過剰な適応が起こりやすい。
  - このことを数学的に整理してみるのが目的。

損失関数と Bias,Variance, Noise

- xが与えられたときの結果：tの推定値＝$y(\mathbf{x})$
- 損失関数: $L(t,y(\boldsymbol{x}))$　　ex. $(y(\mathbf{x})-t)^2$
- 損失の期待値：E[L]を最小化する　t の推定値=E[t|$\mathbf{x}$]
  - この導出は次の次のページを参考にしてください

- E[L]を計算してみると（次のページ参照）

$$\boxed{E[L]=\int (y(\mathbf{x})-E[t\mid \mathbf{x}])^2 p(\mathbf{x})\mathrm{d}\mathbf{x}+\iint (E[t\mid \mathbf{x}]-t)^2 p(\mathbf{x},t)\mathrm{d}t\mathrm{d}\mathbf{x}}$$

  - 第1項は予測値と学習データからの期待値の差の2乗、第2項は雑音(noise)

参考：E[L]の計算

$$L=\left(y(\mathbf{x})-t\right)^2=\left(y(\mathbf{x})-E[t\mid\mathbf{x}]+E[t\mid\mathbf{x}]-t\right)^2$$

$$=\left(y(\mathbf{x})-E[t\mid\mathbf{x}]\right)^2+2\left(y(\mathbf{x})-E[t\mid\mathbf{x}]\right)\left(E[t\mid\mathbf{x}]-t\right)+\left(E[t\mid\mathbf{x}]-t\right)^2$$

第２項の1/2倍をtで周辺化する

$$\int\left(y(\mathbf{x})-E[t\mid\mathbf{x}]\right)\left(E[t\mid\mathbf{x}]-t\right)p(\mathbf{x},t)\,\mathrm{d}t$$

$y(\mathbf{x})-E[t\mid\mathbf{x}]$は$t$の関数ではないので

$$=\left(y(\mathbf{x})-E[t\mid\mathbf{x}]\right)\int\left(E[t\mid\mathbf{x}]-t\right)p(\mathbf{x},t)\mathrm{d}t$$

$$=\left(y(\mathbf{x})-E[t\mid\mathbf{x}]\right)\left\{E[t\mid\mathbf{x}]\int p(\mathbf{x},t)\mathrm{d}t-\int tp(\mathbf{x},t)\mathrm{d}t\right\}$$

$$=\left(y(\mathbf{x})-E[t\mid\mathbf{x}]\right)\left\{E[t\mid\mathbf{x}]p(\mathbf{x})-\int t\frac{p(\mathbf{x},t)}{p(\mathbf{x})}\mathrm{d}t\,p(\mathbf{x})\right\}$$

$$=\left(y(\mathbf{x})-E[t\mid\mathbf{x}]\right)\left(E[t\mid\mathbf{x}]p(\mathbf{x})-E[t\mid\mathbf{x}]p(\mathbf{x})\right)=0$$ よって

$$E\left[\left(y(\mathbf{x})-t\right)^2\right]=\int\int\left(y(\mathbf{x})-E[t\mid\mathbf{x}]\right)^2p(\mathbf{x},t)\mathrm{d}t\mathrm{d}\mathbf{x}$$

$$=\int\left(y(\mathbf{x})-E[t\mid\mathbf{x}]\right)^2p(\mathbf{x})\mathrm{d}x+\int\int\left(E[t\mid\mathbf{x}]-t\right)^2p(\mathbf{x},t)\mathrm{d}t\mathrm{d}\mathbf{x}$$

参考:E[L]を最小化するt の推定値=E[t|$\mathbf{x}$]の導出

$$E[L]=\iint L(y(\mathbf{x}),t)p(\mathbf{x},t)\mathrm{d}t\mathrm{d}\mathbf{x}=\iint(y(\mathbf{x})-t)^2 p(\mathbf{x},t)\mathrm{d}t\mathrm{d}\mathbf{x}$$

$E[L]$を最小化する関数　$y(\mathbf{x})$を求めるには変分法。

この場合は簡単で$E[L]$を$y(\mathbf{x})$で変分（微分）し０とおけばよい

ただし、$\mathbf{x}$は微分の対象ではないので、定数とみなしておくから

$$\frac{\delta E[L]}{\delta y(\mathbf{x})}=\int\frac{\partial}{\partial y(\mathbf{x})}(y(\mathbf{x})-t)^2 p(\mathbf{x},t)\mathrm{d}t=2\int(y(\mathbf{x})-t)p(\mathbf{x},t)\mathrm{d}t=0$$

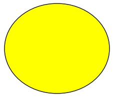
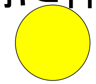
$$\Rightarrow\quad y(\mathbf{x})\int p(\mathbf{x},t)\mathrm{d}t=y(\mathbf{x})p(\mathbf{x})\quad=\quad\int tp(\mathbf{x},t)\mathrm{d}t$$

$$\Rightarrow\quad y(\mathbf{x})=\frac{\int tp(\mathbf{x},t)\mathrm{d}t}{p(\mathbf{x})}=\int t\frac{p(\mathbf{x},t)}{p(\mathbf{x})}\mathrm{d}t=\int tp(t\mid\mathbf{x})\mathrm{d}t=E[t\mid\mathbf{x}]$$

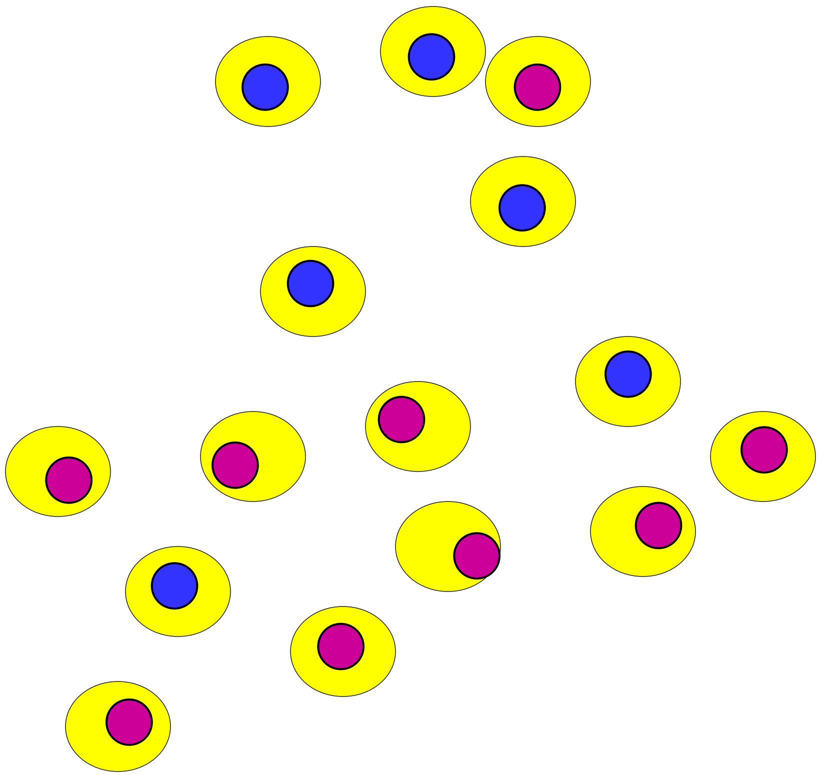
- E[t|$\boldsymbol{x}$]は$\boldsymbol{x}$によって決まる。E[L]は次式でした。

$$\boxed{E[L]=\int (y(\mathbf{x})-E[t\mid\mathbf{x}])^2 p(\mathbf{x})\mathrm{d}\mathbf{x}+\iint (E[t\mid\mathbf{x}]-t)^2 p(\mathbf{x},t)\mathrm{dtd}\mathbf{x}}$$

- 第２項
  - （）内の左の項は、観測値として与えられたxに対して$E[L]$を最小化するtの予測値だから、（）内の右の項すなわち真のt　との差は、観測における誤差と考えられる。
  - $y(\boldsymbol{x})$の作り方で解決できないノイズ

 は、データ点の観測に伴う誤差あるいはノイズの効果を示し、真のデータ点は、大体  のような範囲にある。このノイズの項が既に述べた次の式:

$$\iint (E[t \mid \mathbf{x}] - t)^2 p(\mathbf{x}, t) \mathrm{d}t \mathrm{d}\mathbf{x}$$



$$E[L]=\int(y(\mathbf{x})-E[t\mid\mathbf{x}])^2 p(\mathbf{x})\mathrm{d}\mathbf{x}+\iint(E[t\mid\mathbf{x}]-t)^2 p(\mathbf{x},t)\mathrm{d}t\mathrm{d}\mathbf{x}$$

- $E[L]$の第1項と教師データ集合：Dから機械学習で得た$y(\mathbf{x};D)$の関係について考えてみよう。
- 母集団のモデルとして$p(x,t)$を想定する。このモデルからDという教師データ集合が繰り返し取り出される状況を考えてみる。
- Dからの機械学習の結果の$y(\mathbf{x};D)$の統計的性質は、同じサイズのDを多数回、母集団モデル$p(t,\mathrm{x})$から取り出して、その上で期待値をとった$E_{\mathrm{D}}[y(\mathbf{x};D)]$によって評価する。
- $E[L]$の第1項は$y(\mathbf{x}:D)$とtの最適予測$\mathrm{E}[\mathrm{t}|\mathrm{x}:D]$を用いると次の式

$$\int E_D\left[(y(\mathbf{x}:D)-E[t\mid\mathbf{x}:D])^2\right]p(\mathbf{x})\mathrm{d}\mathbf{x}$$

$$(y(\mathbf{x}:D)-E[t\,|\,\mathbf{x}:D])^2=(y(\mathbf{x}:D)-E_D[y(\mathbf{x}:D)]+E_D[y(\mathbf{x}:D)]-E[t\,|\,\mathbf{x}:D])^2$$
$$=(y(\mathbf{x}:D)-E_D[y(\mathbf{x}:D)])^2+(E_D[y(\mathbf{x}:D)]-E[t\,|\,\mathbf{x}:D])^2$$
$$+2(y(\mathbf{x}:D)-E_D[y(\mathbf{x}:D)])(E_D[y(\mathbf{x}:D)]-E[t\,|\,\mathbf{x}:D])$$

この式をE_D[]すると、第3項は消え

$$E_D[(y(\mathbf{x}:D)-E[t\,|\,\mathbf{x}:D])^2]$$
$$=\underbrace{E_D[(y(\mathbf{x}:D)-E_D[y(\mathbf{x}:D)])^2]}_{\text{第1項はvariance}}+\underbrace{E_D[(E_D[y(\mathbf{x}:D)]-E[t\,|\,\mathbf{x}:D])^2]}_{\text{第2項は}\mathrm{bias}^2}$$

- variance: $y(\mathbf{x})$の機械学習による推定値が、教師データ集合によって変動する度合いの期待値:教師データに依存しすぎるモデルになって新規データの予測誤差が悪化する度合い

- bias^2: $y(\mathbf{x})$の機械学習による推定値が、損失の期待値:$\mathrm{E}[L]$を最小化するtからずれる度合いの期待値:モデルを記述が単純になるとき予測誤差が悪化する度合い。

以上により損失の期待値：E[L]=bias²+variance+noise

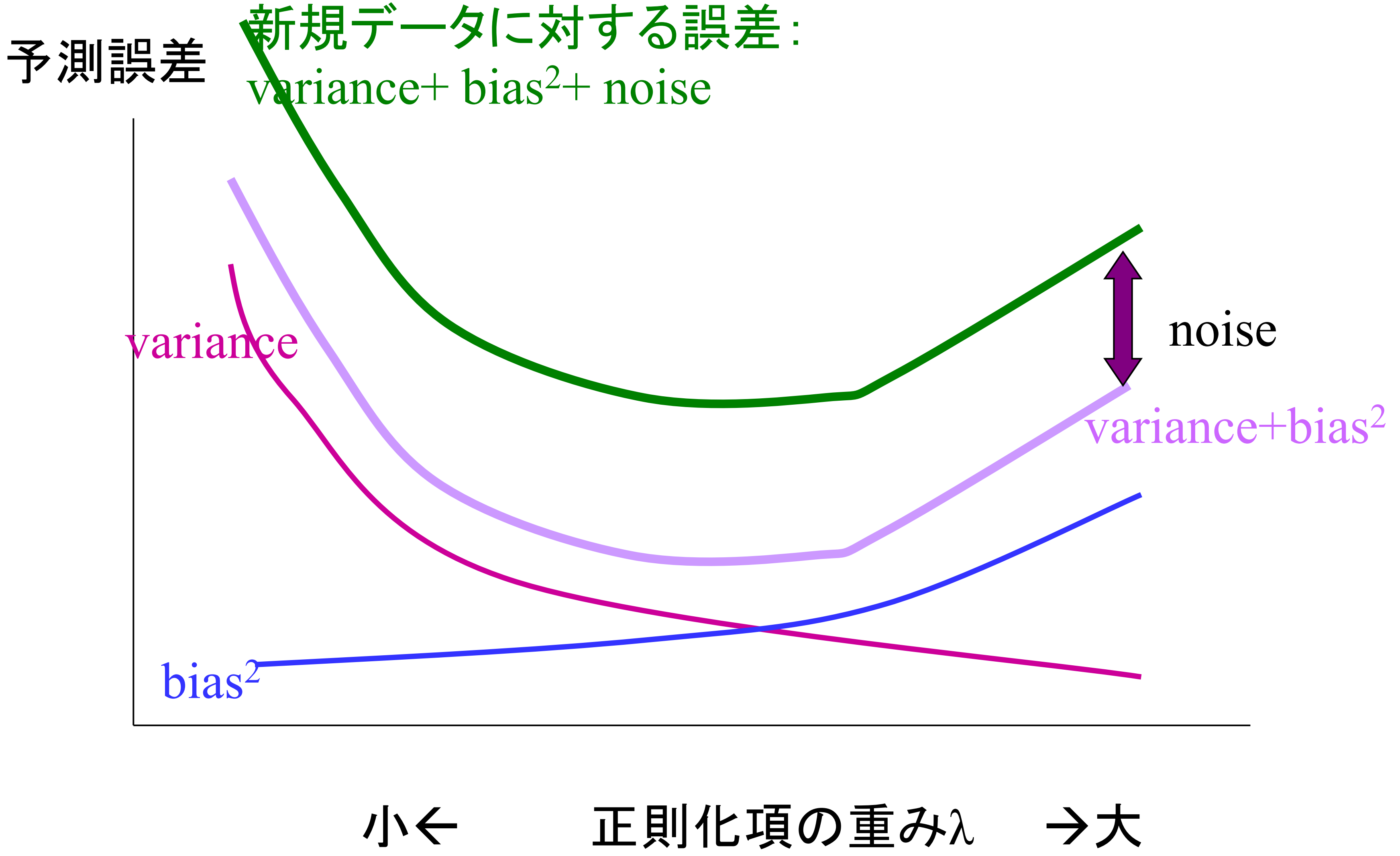
$$E[L] =$$

$$\int (E_D[y(\mathbf{x}:D)] - E[t \mid \mathbf{x}:D])^2 p(\mathbf{x})\mathrm{d}\mathbf{x} \qquad \text{bias}^2$$

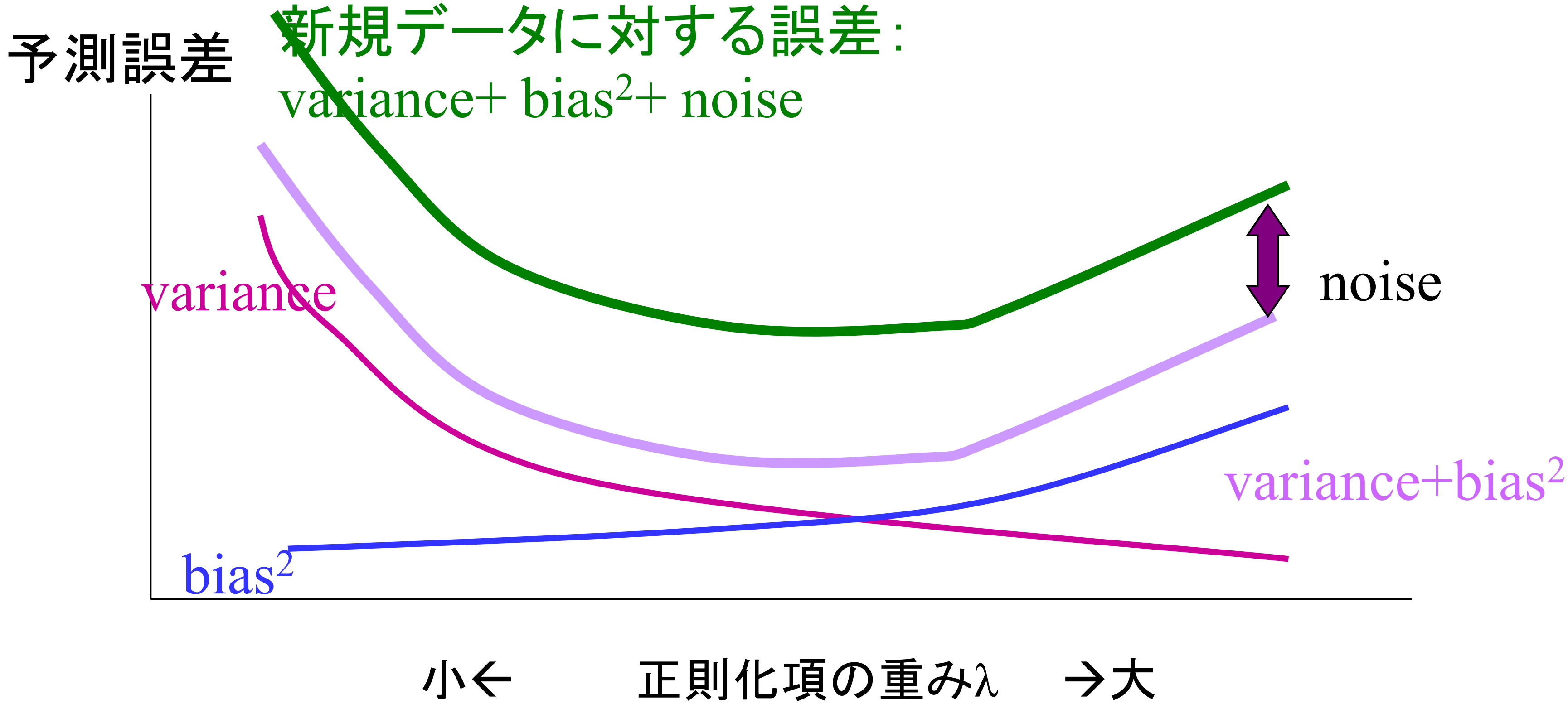
$$+ \int E_D[(y(\mathbf{x}:D) - E_D[y(\mathbf{x}:D)])^2] p(\mathbf{x})\mathrm{d}\mathbf{x} \qquad \text{variance}$$

$$+ \iint (E[t \mid \mathbf{x}:D] - t)^2 p(\mathbf{x}:D, t)\mathrm{d}\mathbf{x}\mathrm{d}t \qquad \text{noise}$$

bias^2とvarianceの間には次のページに示すようなトレードオフがある。



予測誤差
新規データに対する誤差：
variance+ bias2+ noise
noise
variance
variance+bias2
bias2
小← 正則化項の重みλ →大



- L2正則化の場合

観測データに大きく異存←小　λ　大→正則化項(**事前分布**)に大きく依存

- L1正則化の場合：重みがゼロ化される次元をみると

ゼロの次元が少なく複雑 ←小　λ　大→ゼロの次元が多く単純

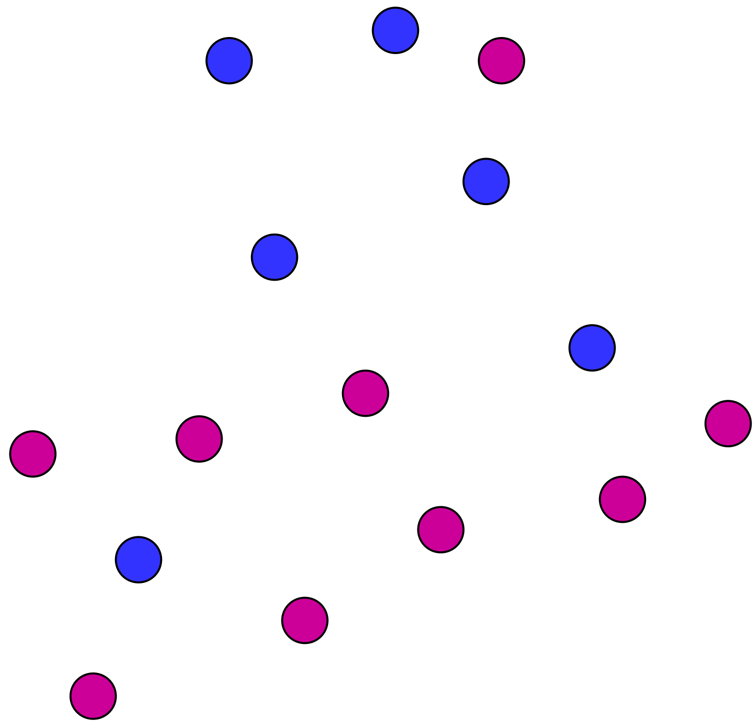
bias^{2}とvarianceの間のトレードオフをK-Nearest Neighbor法と線形回帰で具体的に見てみよう。

K-Nearest Neighbor法

- 2クラスへの分類問題で考える。
- 教師データはクラス: ● とクラス: ● と判定された相当数があるとする。
- 未知のデータxがクラス ●／● である確率は
  - xに近いほうからK個の教師データ点のうちでクラス●／●であるものの割合

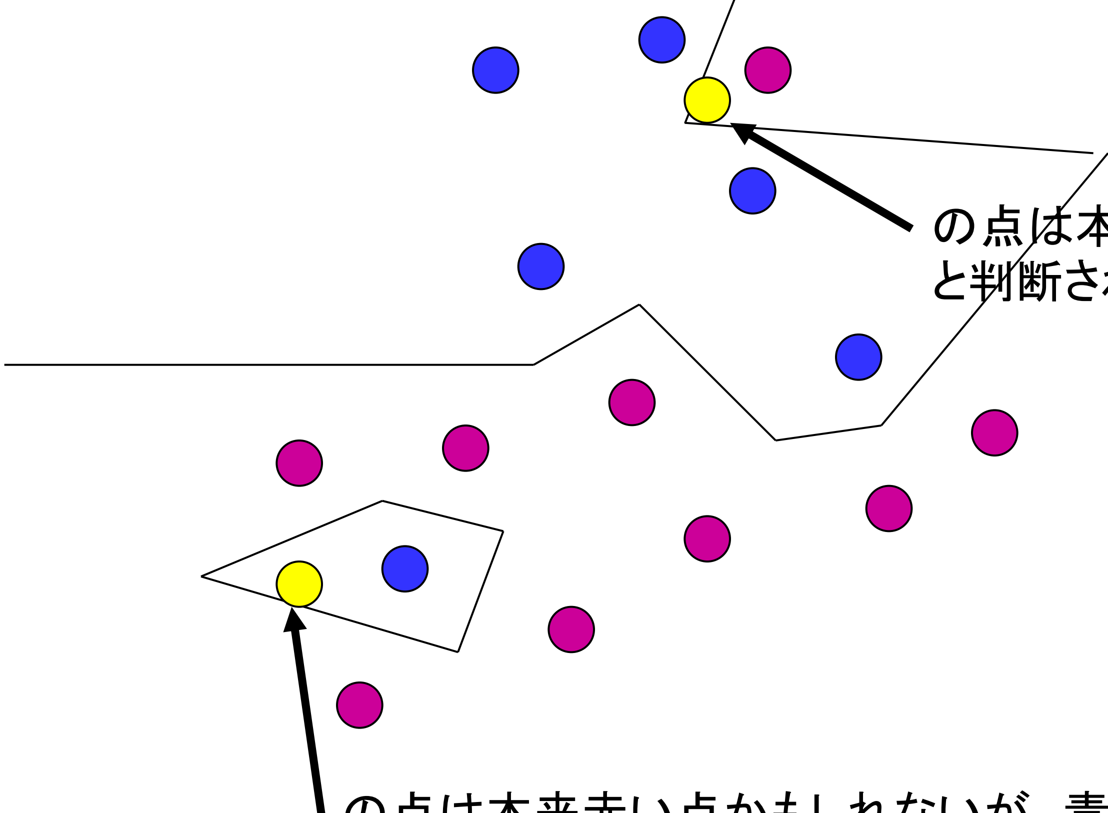
✓至ってシンプルだがかなり強力。

下の図のような教師データの配置で考える



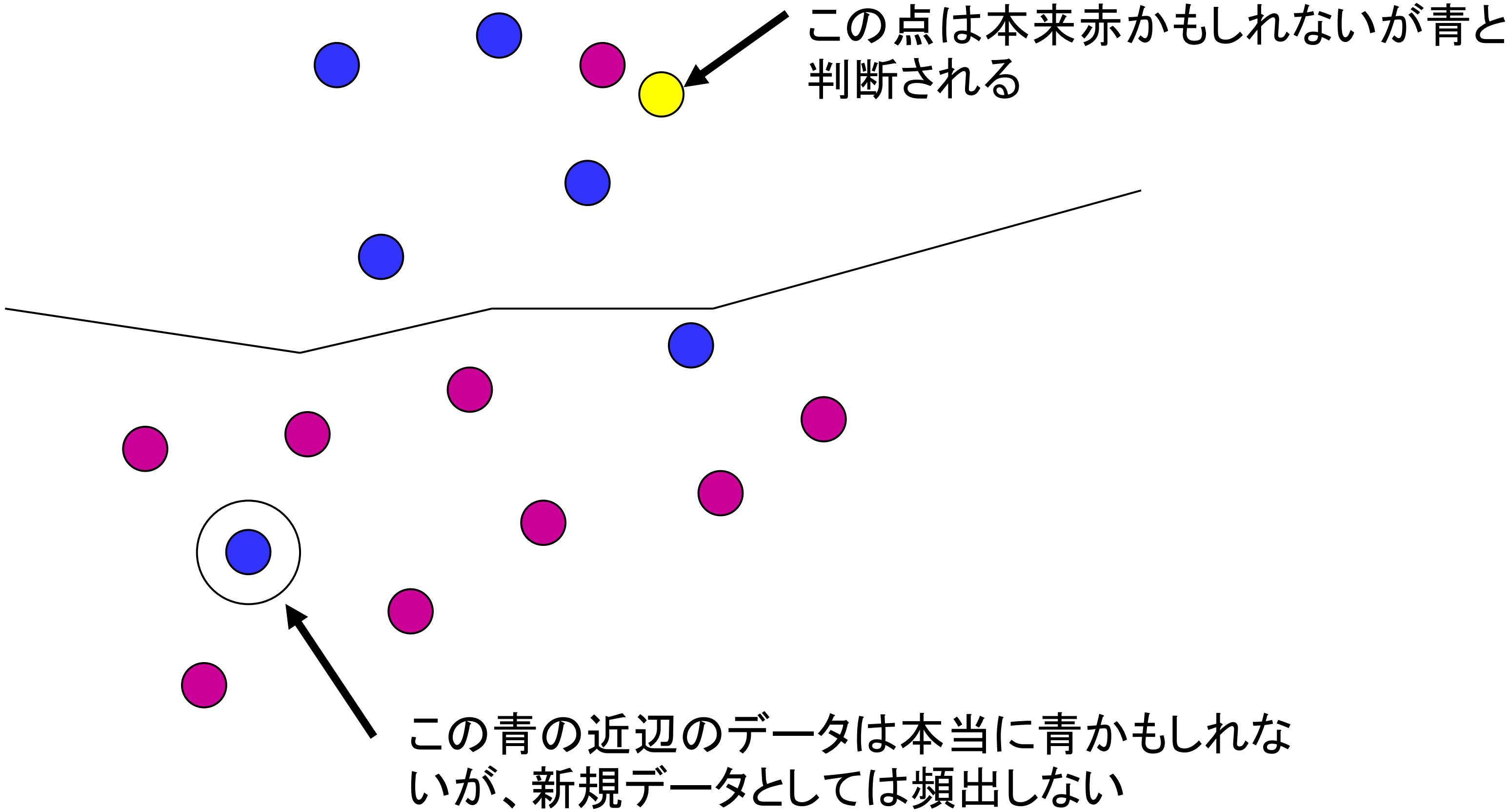
K=1の場合：クラス青，赤の確率が等しい境界線は以下のようにかなり複雑。相当多くのパラメターを使わないと記述できない。教師データ数に強く依存。

は新規に到着した分類すべきデータ

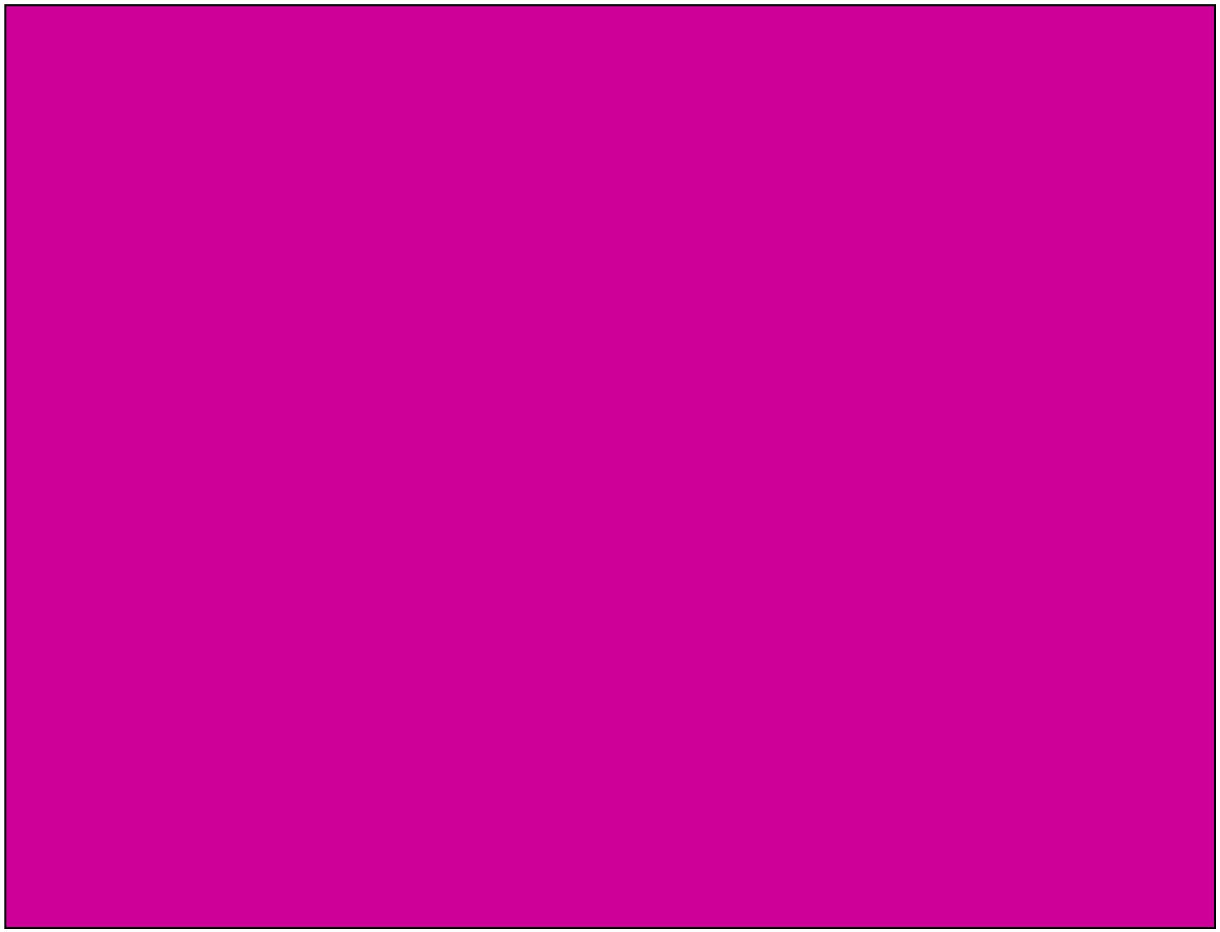


K=3の場合のクラス間の境界

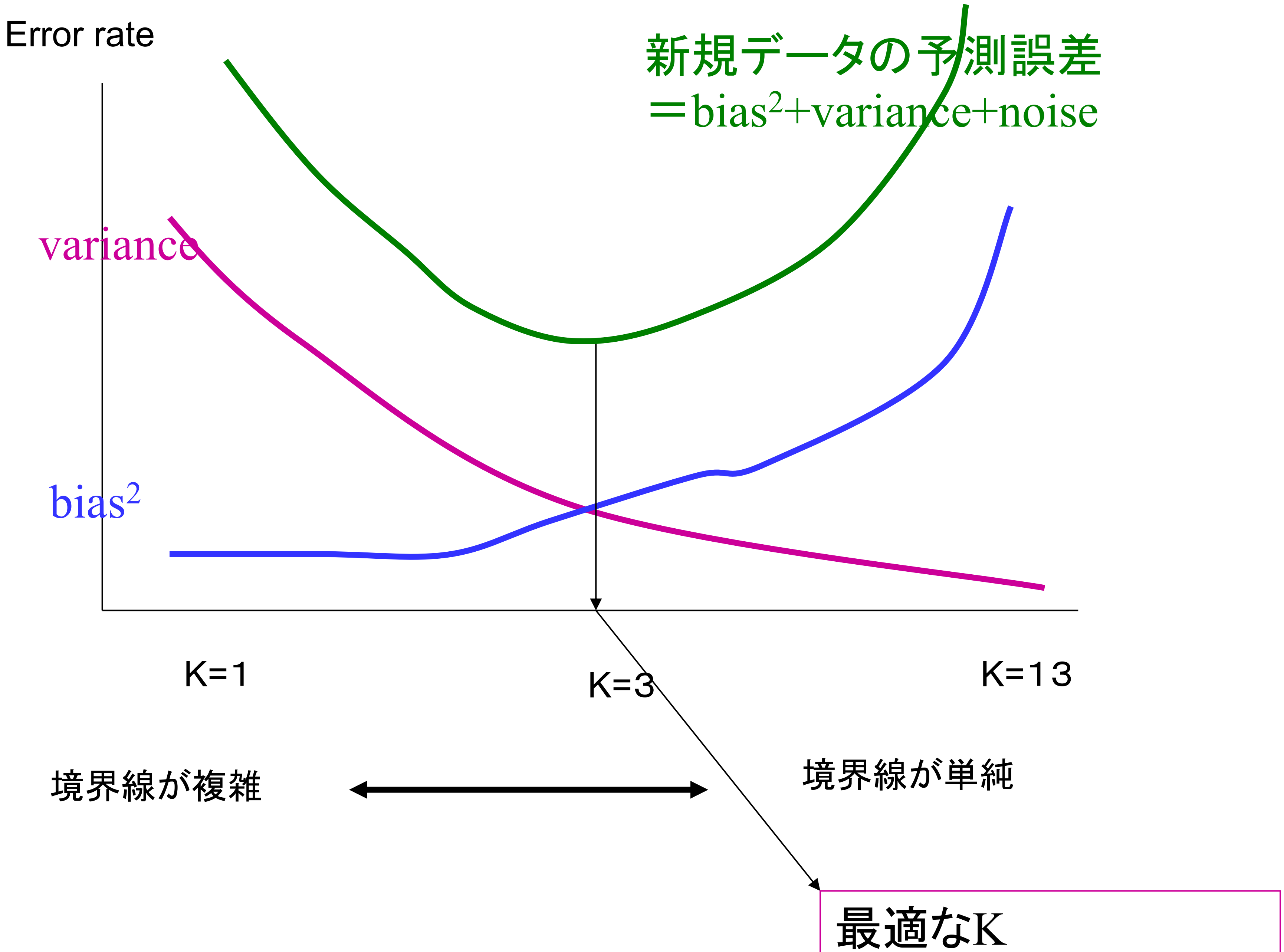
境界線はだいぶ滑らか。K=1の場合より境界を決めるパラメターは多い



K=13以上だと、どんな新規データでも赤と判定される。



- K=1だと非常に複雑な境界線であり、個々の教師データに強く依存した結果をだすため、過学習をしやすい。 varianceが大きい。
- Kが大きくなると、境界線は平滑化される方向に進む。教師データを適当な数使って結果を出すので、過学習を起こしにくい。
- Kが非常に大きくなると、境界線はますます滑らか（＝いい加減？）になり、あるところから個別の教師データの影響が無視され、モデルとして大域のデータに依存し、個別データに対する精密さを欠くため、新規データを正確に分類できなくなってくる。$bias^2$ が大きい。
- 以上のから、**$bias^2$とvariance**の間には次ページの図のような関係が見てとれる。


Error rate
新規データの予測誤差
＝bias²+variance+noise
variance
bias²
K=1
K=3
K=13
境界線が複雑
境界線が単純
最適なK

bias^{2}とvarianceの間のトレードオフを線形回帰で具体的に見てみよう。

まず線形モデルのパラメーター**w**推定の復習から

$$y = \langle \mathbf{x}, \mathbf{w} \rangle + \varepsilon = \sum_{i=0}^{K} w_i x_i + \varepsilon$$

ただし、$\mathbf{x} = (1, x_1, \ldots, x_K)^T, \mathbf{w} = (w_0, w_1, \ldots, w_K)^{\mathrm{T}}$

εはノイズで$N(0, \sigma^2)$と考える。

➢入力ベクトル : $\mathbf{x}$ から出力 : y を得る関数がxの線形関数（wとxの内積）にノイズが加算された場合を再掲

$$y = \langle \mathbf{x}, \mathbf{w} \rangle + \varepsilon = \sum_{i=0}^{K} w_i x_i + \varepsilon \qquad ただし、\mathbf{x} = (1, x_1, \ldots, x_K)^T, \mathbf{w} = (w_0, w_1, , \ldots, w_K)^{\mathrm{T}}$$

εはノイズで$N(0, \sigma^2)$と考える。

➢得られたN個の観測データ の組（$\mathbf{y}$,$\mathbf{X}$）に対して2乗誤差を最小化するようにwを推定し$\hat{\mathbf{w}}$ を得る。

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1{}^T \\ \vdots \\ \mathbf{x}_N{}^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{NK} \end{pmatrix} \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{pmatrix}$$

$\varepsilon_i \ (i = 1, \ldots, N)$は$N(0, \sigma^2)$の$iid$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} = \mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon} \quad \Rightarrow \quad \hat{\mathbf{w}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \qquad (0)$$

- ここで、前にやった損失の期待値　E(L) を思いだそう
- ただし、新規の未知データ$y_0,\mathbf{x}_0$は以下の通り

$$E_{y_0\mathbf{x}_0}[L]=\int (y(\mathbf{x}_0)-E_{y_0}[y(\mathbf{x}_0)\mid \mathbf{x}_0])^2 p(\mathbf{x}_0)\mathrm{d}x_0$$

$$+\iint (E_{y_0}[y(\mathbf{x}_0)\mid \mathbf{x}_0]-y_0)^2 p(\mathbf{x}_0,y_0)\mathrm{d}y_0\mathrm{d}\mathbf{x}_0 \quad (loss0)$$

$$y_0=\langle \mathbf{x}_0,\mathbf{w}\rangle+\varepsilon \qquad (loss1)$$

$$E_{y_0}[y(\mathbf{x}_0)\mid \mathbf{x}_0]=\int y_0 p(y_0\mid \mathbf{x}_0)\mathrm{d}y_0$$ だったが、$(loss1)$を使うと

$$E_{y_0}[y_0\mid \mathbf{x}_0]=\int (\langle \mathbf{x}_0,\mathbf{w}\rangle+\varepsilon)p(y_0\mid \mathbf{x}_0)\mathrm{d}y_0$$

$$=\int \langle \mathbf{x}_0,\mathbf{w}\rangle p(y_0\mid \mathbf{x}_0)\mathrm{d}y_0+\int \varepsilon p(y_0\mid \mathbf{x}_0)\mathrm{d}y_0=\langle \mathbf{x}_0,\mathbf{w}\rangle \quad (loss2)$$

$$E_{y_0\mathbf{x}_0,D}[L]=\int\int (y(\mathbf{x}_0)-\langle \mathbf{x}_0,\mathbf{w}\rangle)^2 p(\mathbf{x}_0,\mathbf{y})\mathrm{d}\mathbf{x}_0\mathrm{d}\mathbf{y}+\int\iint (\langle \mathbf{x}_0,\mathbf{w}\rangle-y_0)^2 p(\mathbf{x}_0,y_0,\mathbf{y})\mathrm{d}\mathbf{x}_0\mathrm{d}\mathrm{y}_0\mathrm{d}\mathbf{y}$$

第2項　$\int\iint (\langle \mathbf{x}_0,\mathbf{w}\rangle-y_0)^2 p(\mathbf{x}_0,y_0,\mathbf{y})\mathrm{d}\mathbf{x}_0\mathrm{d}\mathrm{y}_0\mathrm{d}\mathbf{y}$

$$=\int\iint (\langle \mathbf{x}_0,\mathbf{w}\rangle-(\langle \mathbf{x}_0,\mathbf{w}\rangle+\varepsilon))^2 p(\mathbf{x}_0,y_0,\mathbf{y})\mathrm{d}\mathbf{x}_0\mathrm{d}\mathrm{y}_0\mathrm{d}\mathbf{y}$$

$$=\iint\int \varepsilon^2 p(\mathbf{x}_0,y_0,\mathbf{y})\mathrm{d}\mathbf{x}_0\mathrm{d}\mathrm{y}_0\mathrm{d}\mathbf{y}=\sigma^2$$ $\Rightarrow$ 新規の未知データの観測に伴う雑音

- 次は$E_{y_0\mathbf{x}_0,D}[L]$の第1項$=\iiint\left(y(\mathbf{x}_0)-\langle\mathbf{x}_0,\mathbf{w}\rangle\right)^2 p(\mathbf{x}_0,y_0,\mathbf{y})\mathrm{d}\mathbf{x}_0\mathrm{d}y_0\mathrm{d}\mathbf{y}$
  すなわち観測データ（あるいは計画行列）（**y**,**X**）＝Dを多数作って学習データとする部分について考える。

- Xに対して繰り返しyを観測することでDを動かした場合の期待 値：$\mathrm{E_D}[..]$を求めてみよう。

- 重みwの期待値：$\hat{\mathbf{w}}$ のD動かした場合の期待値 $=E_D[\hat{\mathbf{w}}]$

$$\boxed{E_D[\hat{\mathbf{w}}]=E_D\left[\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}\right]=E_D\left[\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{w}+\boldsymbol{\varepsilon})\right]=\mathbf{w}}$$

レポート課題1：共分散行列を求めよ

$$\mathrm{cov}_D[\hat{\mathbf{w}}]=?$$

XはDにおいては定数なので、$(\mathrm{X^TX})^{-1}\mathrm{X^T}$も定数と見なせることに注意

$$E_{y_0\mathbf{x}_0,D}[L]\text{の第1項} = \iiint (y(\mathbf{x}_0) - \langle \mathbf{x}_0, \mathbf{w} \rangle)^2 p(\mathbf{x}_0, y_0, \mathbf{y}) \mathrm{d}\mathbf{x}_0 \mathrm{d}y_0 \mathrm{d}\mathbf{y}$$

$$E_{y_0\mathbf{x}_0,D}[(y(\mathbf{x}_0 : D) - \langle \mathbf{x}_0, \mathbf{w} \rangle)^2]$$

$$= E_{y_0\mathbf{x}_0,D}[(y(\mathbf{x}_0 : D) - E_D[y(\mathbf{x}_0 : D)])^2] + E_{y_0\mathbf{x}_0,D}[(E_D[y(\mathbf{x}_0 : D)] - \langle \mathbf{x}_0, \mathbf{w} \rangle)^2]$$

$$-(loss10)$$

$E_D[y(\mathbf{x} : D)]$はDを動かしての期待値だが、$\mathbf{X}$は同一でyの観測だけを繰り返しているので、この期待値は$E_D[\hat{\mathbf{w}}]$になる。

$$\Rightarrow E_D[y(x_0 : D)] = \langle \mathbf{x}_0, E_D[\hat{\mathbf{w}}] \rangle = \langle \mathbf{x}_0, \mathbf{w} \rangle$$

$y(\mathbf{x}_0 : D)$はあるDに対する予測だから、

Dに対する正規方程式の解(0)より　　$\mathbf{x}_0{}^T (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{y}$

$$(loss10) = \text{variance} + \text{bias}^2$$

$$= E_{y_0\mathbf{x}_0,D}[(\mathbf{x}_0{}^T (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{y} - \langle \mathbf{x}_0, \mathbf{w} \rangle)^2] + E_{y_0x_0,D}[(\langle \mathbf{x}_0, \mathbf{w} \rangle - \langle \mathbf{x}_0, \mathbf{w} \rangle)^2]$$

$$\Rightarrow \quad \text{bias}^2 = 0$$

レポート課題2：bias^{2}が0にならない状況を考察せよ

$$E_{y_0\mathbf{x}_0,D}[L]\text{の第1項} = \iiint \left(y(\mathbf{x}_0) - \langle \mathbf{x}_0, \mathbf{w} \rangle\right)^2 p(\mathbf{x}_0, y_0, \mathbf{y})\mathrm{d}\mathbf{x}_0 \mathrm{d}y_0 \mathrm{d}\mathbf{y}$$

$$E_{y_0\mathbf{x}_0,D}[(y(\mathbf{x}_0 : D) - \langle \mathbf{x}_0, \mathbf{w} \rangle)^2]$$

$$= E_{y_0\mathbf{x}_0,D}[(y(\mathbf{x}_0 : D) - E_D[y(\mathbf{x}_0 : D)])^2] + E_{y_0\mathbf{x}_0,D}[(E_D[y(\mathbf{x}_0 : D)] - \langle \mathbf{x}_0, \mathbf{w} \rangle)^2] \quad (loss10)$$

$$\boxed{\text{variance of } (loss10) = E_{y_0\mathbf{x}_0,D}[(\mathbf{x}_0{}^T \left(\mathbf{X}^\mathrm{T}\mathbf{X}\right)^{-1} \mathbf{X}^\mathrm{T}\mathbf{y} - \langle \mathbf{x}_0, \mathbf{w} \rangle)^2]}$$

レポート課題3:variance of (*loss 10*)を求めよ

$\mathbf{X}$は十分大きく多様な説明変数からなり
$\mathbf{X}^\mathrm{T}\mathbf{X} = N \cdot E_{\mathbf{x}_0}\left[\mathbf{x}_0\mathbf{x}_0{}^\mathrm{T}\right]$と近似できるとする。

レポート課題4:この場合variance of (*loss 10*)の近似式を求めよ

過学習：over-fittingと bias^2-variance分解

- bias^2-variance分解は過学習現象を扱う数学的概念として便利

- 教師データによる学習の目的は未知のデータの正確な分類や識別
- 過学習(over-fitting)
  - 学習するモデルを複雑な（＝パラメタ数の多い）ものにすると過学習が起こりやすい。
  - モデルの良さ（＝（対数）尤度あるいは2乗誤差などの損失$^{-1}$）を最大化し、かつ簡単なモデルであるほど良い
  - モデルの簡単さを表すのは線形回帰における正規化項（正則化項とも呼ぶ）。cf.情報量基準（AIC，BIC）、MDL