正規分布　ガウス分布（ガウスのアイディア）

（天体の）x_i X ε_n

観測値 = （真の値）+（誤差）

「誤差のバラツキ」が分かれば「真の値」を推定できる。

‖

ε分布　確率密度 $f(\varepsilon) = f(x_i - X)$

1809 ガウス
太陽の周りを楕円軌道で公転する天体の運行に関する理論

↓ ラプラスへ

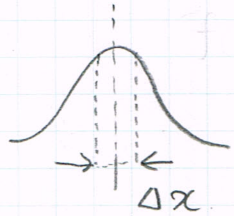
最小二乗法の前提として応用

ガウスさんの経験から3つの公理（axiom）

A1. 絶対値の小さい誤差は絶対値の大きい誤差よりも多く発生

A2. 絶対値が同じ正・負の誤差は同じくらいに発生

A3. 絶対値の（とても）大きい誤差は非常に起こりにくい



$f(x_i - X)\cdot\Delta x$ は $x_i = X$ $(\varepsilon = 0)$ で最大値

観測値全てが Δx内に入る同時確率は

$$f(x_1 - X)\Delta x \times f(x_2 - X)\cdot\Delta x \times \cdots \times f(x_n - X)\cdot\Delta x$$

$$= (\Delta x)^n \prod_{i=1}^{n} f(x_i - X) = \left(\frac{d}{dx}\right)^n \cdot P(X) \text{ とすると}$$

$$P(X) = f(x_1 - X) f(x_2 - X) \cdots f(x_n - X) \quad \cdots Ⓐ$$

Pも明らかに、$x_i = X$ $(i = 1 \sim n)$ で最大値

⇄ $P'(X = \bar{x}) = 0$ と仮定できる。（連続と仮定）…①

Ⓐの自然対数 $\ln P(X) = \sum_{i=1}^{n} \ln(x_i - X)$

Xで微分 $\dfrac{P'(X)}{P(X)} = -\sum_{i=1}^{n} \dfrac{f'(x_i - X)}{f(x_i - X)}$ …Ⓑ

$\ln(x) = \frac{1}{x}$ と $\frac{d}{dx}(f(g(x))) = f'(g(x))\cdot g'(x)$ より

正規分布　ガウス分布…ガウスのアイディア

Ⓑと $P'(X=\bar{x})=0$ より

$$\sum_{i=1}^{n}\frac{f'(x_i-\bar{x})}{f(x_i-\bar{x})}=0 \quad \text{また} \quad \sum_{i=1}^{n}(x_i-\bar{x})=0$$

であることから、

$$\frac{f'(x)}{f(x)}=ax$$ （注、証明 → 別紙$\boxed{3}$）

$$\rightleftarrows \quad f(x)=A\cdot e^{\frac{ax^2}{2}}$$

A3より、$a<0$ なので $a=-\frac{1}{\sigma^2}$ と置くと、

$$f(x)=A\exp\left\{-\frac{1}{2}\cdot\left(\frac{x}{\sigma}\right)^2\right\}$$

全確率が1となるので

$$\int_{-\infty}^{\infty}f(x)dx=1=\int_{-\infty}^{\infty}A\exp\left\{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2\right\}$$

$$=A\sqrt{2\pi\sigma^2}$$

cf.
（ガウスの公式）

$$\int_{-\infty}^{\infty}e^{-ax^2}dx=\sqrt{\frac{\pi}{a}}$$

$(a>0)$

$$\therefore A=\frac{1}{\sqrt{2\pi\sigma^2}}$$

$$\rightleftarrows \quad f(\varepsilon)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2}\left(\frac{\varepsilon}{\sigma}\right)^2\right\}$$

$$f(x-X)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2}\left(\frac{x-X}{\sigma}\right)^2\right\}$$

注）ただ、$P'(X=\bar{x})\neq 0$ という仮定は公理にはない。

$n(\geq 3)$として、 $\sum_{i=1}^{n} x_i = 0$ ……① を満足するように x_i が変化するとき、

関数 $h(x)$ $\sum_{i=1}^{n} h(x_i) = 0$ ……② が常に成立しているならば、

$h(x) = kx$ （kは任意の定数）である。

証明

$$x_n = -\sum_{i=1}^{n-1} x_i \quad \cdots ①'$$

$$\frac{\partial}{\partial x_1} h(x_1) + \frac{\partial h(x_n)}{\partial x_n} \cdot \frac{\partial x_n}{\partial x_1} = 0$$

$$h'(x_1) + h'(x_n) \cdot \frac{\partial x_n}{\partial x_1} = 0$$

①′から、$\dfrac{\partial x_n}{\partial x_1} = -1$ より、 $h'(x_1) = h'(x_n)$

同様に $h'(x_1) = h'(x_2) = \cdots = h'(x_n)$

$h'(x_i)$は xによらない事から、 $h'(x) = k$（定数）

$h(x) = kx + c$、

②から、 $c = 0$ 　　なので、 $h(x) = kx$、 //

分散・共分散行列　と　機械学習
ML

ノート
cf. 多変数の正規分布

↑(1変数の)確率の分散の考え方を多変数・多次元に拡張したもの

1変数＝分散　$V(X) = \sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad (= \sigma_{xx})$

↑平均からの距離^{2}の総和
データの回転質量(2次モーメント)
$\|$

分散を最小
↓↑
最小二乗法

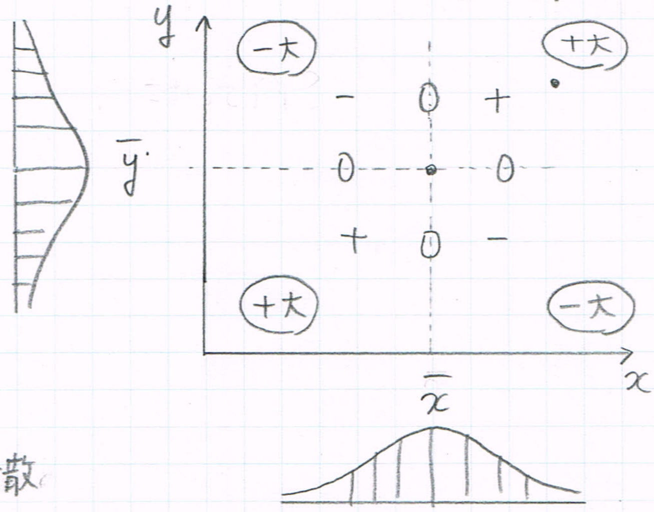
MLっぽい

2変数間＝共分散

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

これって何？

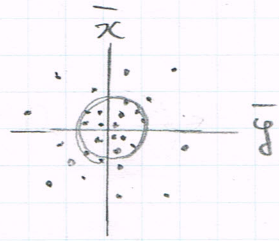
$(x_i - \bar{x})(y_i - \bar{y})$ は



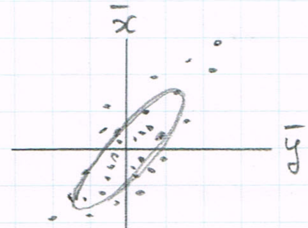
← このようになる。
⇓
全てのデータについての総和

共分散
σ_{xy} の特徴

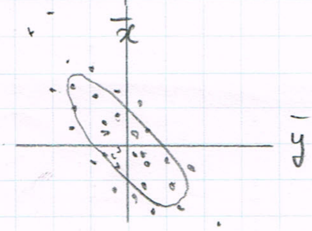
① 全体の(x, y)に関連性がなければ、
＋点と－点が打ちけしあい、0になる。



② x増 → y増 の関係が多ければ、
＋点が多くなり、「＋」になる



③ x増 → y減の関係が多ければ
－点が多くなり、「－」になる。



「分散・共分散行列」と 機械学習
ML

分散共分散行列の定義

$\sigma_{xx} = (\sigma_x)^2$, $\sigma_{xy} = \sigma_{yx}$

2変数

$$\Sigma = \Sigma_{xy} = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{pmatrix}$$

縮約記法的…

3変数

$$\Sigma = \Sigma_{xyz} = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{xy} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{xz} & \sigma_{yz} & \sigma_{zz} \end{pmatrix}$$

「ことば」

対角成分は「分散」 それ以外は「共分散」の「行列」

例 xとyには正の相関があるが
zは独立しているデータセットのΣ →
$$\begin{pmatrix} + & + & 0 \\ + & + & 0 \\ 0 & 0 & + \end{pmatrix}$$

★ N変数の中の任意の2変数の共分散($_NC_2$個)とN変数の(N個)の分散が、対称な正方行列中に整理されて入っている。

☆ N次元ベクトルに拡張した時の「分散」といえる。

--→ 単に「分散(行列)」と言う人もいる
(共分散はビルトイン)

↓ (ベイズ勉強会の Σ)

$$f(\vec{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{|\Sigma|}} \exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})\right\}$$

多変数(n次元ベクトル)$\vec{x}$の正規分布もΣを使って美しく表せる。 → cf. ノート 多変数の正規分布

分散・共分散行列 と 機械学習
ML

標準化

そもそもの数値のスケールが異なるデータ間の比較

- 体重，身長 ← 単位すらちがう
- 国語と算数の点数 ← 同じ国語の点数でも内容がちがえば…

→ $\dfrac{x-\bar{\mu}}{\delta}$

標準化する事で異なるスケールをあわせる。
⋮
例）偏差値

↓ データを標準化すると、

分散 ⇒ 1，共分散 ⇒ 相関係数 になる。

分散 共分散行列 ⇒ 相関行列 になる。
‖
Kaggleで良く出てくる Correlations ヒートマップ

python : pandas_profiling.ProfileReport (df)

の Pearson の図

Peason → パラメトリック検定相関
Spearman → データ順位相関

有意水準以上である
必要性がある。(正規分布仮定が可能)

ところで 相関係数って… 計算してみると、

$$r_{xy} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2}\sqrt{\Sigma(y_i-\bar{y})^2}}$$

$$= \frac{\vec{V_x}\cdot\vec{V_y}}{|\vec{V_x}||\vec{V_y}|} = \cos\theta$$

コサイン類似度 → ML 自然言語解析 画像解析

なので、 $-1 \leqq r_{xy} \leqq 1$ です。

相関行列の成分の，[−1，1] を色域に変換すれば、

多変数間のコサイン類似度を「1見して分かる絵」にしてくれます。 ← Kaggle