機械学習(2) 重回帰

情報科学類　佐久間　淳

再掲

機械学習の流れ

データ

新元号「令和」のもとで初めてとなる予算編成に向けて国の財政制度等審議会は、予算規模の拡大が続く中、社会保障費などの歳出の抑制策を検討する新たな部会を設け、集中的に議論を行うことになりました。
財務大臣に予算の在り方などを提言する財政制度等審議会は、４日から「令和」のもとで初めてとなる、来年度の予算編成に向けて議論を始めました。

https://unsplash.com/photos/IbPxGLgJiMI
Erik-Jan Leusink

NHK NEWS WEB

特徴量(データのベクトル表現)

$$X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\},$$
$$\boldsymbol{x}_i \in \mathbb{R}^D$$

機械学習(教師つき, 教師なし)
特徴量から概念への写像

概念の獲得
スパム、ネコ, 政治ニュース, etc.

再掲

線形モデル

- 目標値と特徴量の線形関係 $t = wx + b$
  - 例：体重xから寿命tを予測する
  - 例：年齢xから総資産額tを予測する
- パラメータ w
  - 特徴量xと目標値tの関係の強さ (絶対値が大きいほど強い関係)
  - 符号が正：特徴量が大きい値をとるほど、目標値も大きい
    - 例：年齢が増えれば総資産額は大きくなる
  - 符号が負：特徴量が小さい値をとるほど、目標値も大きい
    - 例：体重が増えれば寿命は少なくなる
- バイアス b
  - 特徴量が0の時の予測値

再掲

平均二乗誤差

- 誤差：実際の目標値と予測した目標値の差

$$\text{error} = |t_i - (wx + b)|$$

- モデル $wx + b$ における平均二乗誤差
  - 全訓練事例に対する二乗誤差の平均

$$E(w) = \frac{1}{N}\sum_{i=1}^{N}(t_i - (wx_i + b))^2$$

- よいモデル＝平均二乗誤差を最小化するモデル
- 最良のモデルを求めるには？
- 以下を最小にするようなw,bを求める

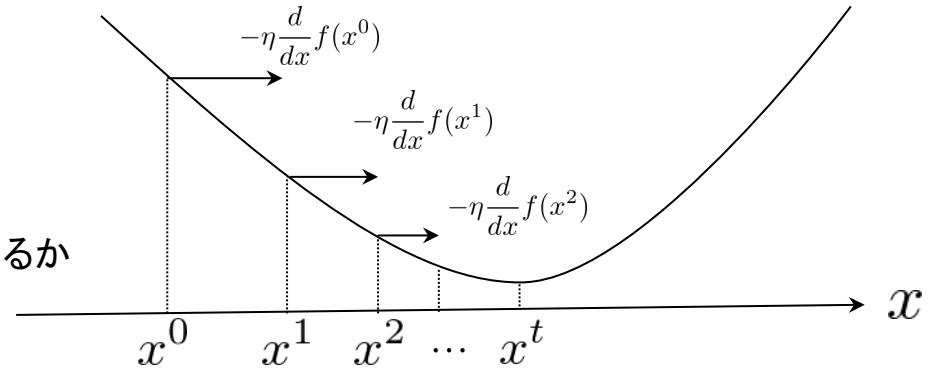
$$\min_{w,b} \frac{1}{N}\sum_{i=1}^{N}(t_i - (wx_i + b))^2$$ …どうやって？

再掲

最急降下法

- 問題 $\min_x f(x)$ となるxを求める
  - 関数fは微分可能
- アルゴリズム
  1. $t = 0, x^0$をランダムに初期化
  2. $x^{t+1} \leftarrow x^t - \eta \frac{d}{dx} f(x^t)$
  3. $t \leftarrow t + 1$
  4. $|x^{t+1} - x^t| < \epsilon$

ステップサイズパラメータη：1ステップにどの程度更新させるか
収束判定パラメータ ϵ



質問・コメント（特徴量）

- 駐車場の有無はどの属性ですか? あるいは無理に駐車場の数にして数値にしたがいいですか?
  - 有無はtrue/falseの離散属性。駐車場の数は連続属性として扱えます。駐車場の数のほうが情報量が多いので属性として優れていますが、調査に手間がかかります
- カテゴリカル/数値/順序の属性がうまく分類できなかったのですが何か分かりやすい線引きはありませんか?、連続値と離散値の例をもっと知りたいと思った
  - 連続性と順序があるかどうかがポイントです。りんご・バナナ・みかん、の間に連続性はありませんが、バナナのサイズS,M,Lには順序があります。バナナの長さ(e.g. 15cm)には連続性があります
- 方角は数値属性?
  - 興味深い質問です。方角を角度(radian)で表示すれば、連続性があるので連続属性に見えますが、2π radと0 radは同じ値なので、通常の実数とは連続性の性質が異なります。このような方向を扱うための統計としてdirectional statisticsなどの学問があります
  - 方角をたとえば東西南北、とすれば、離散属性として扱えます
  - 不動産の価格評価の場合、人気に合わせて南向き>東向き>西向き>北向き、という順序属性とすることもできます

質問・コメント（特徴量）

- データの残渣はどのように消去されるべきなのか?
  - 残渣というのがなにかわかりませんが、欠損値(かけている値)などある場合には、平均値で埋めたりします
- 味のみの特徴でwineの人気予測が出来ると思えない。他に色味、香り、ラベリングデザインなどがあると思われる。
  - 色味、香り、などを数量として扱うために、特徴量としてつかった化学的・物理的な評価量を評価していると考えられます
  - ラベリングデザインなど、人間の感覚でしか理解できないものをどのように機械学習に取り込んでいくか、これは授業終盤頃扱うことができると思います

質問・コメント（モデルと誤差）

- バイアスを日本語で説明するとどのようになるのか?
  - 特徴量にゼロ(ゼロベクトル)を入力した時に出力される値のことです
- 線形モデルの定数項と未学習のときの関数をどちらもバイアスと呼ぶのかがわかりません。
  - 未学習でも、学習済みでも、定義としての線形モデルでも、定数項の部分をバイアスと呼びます
- 平均二乗誤差などで4乗以上ではなく2乗にしているのは計算の複雑さ以外にも理由がありますか?
  - あります。詳しくは黒板で
- 絶対値の総和で誤差を計算したほうが二乗誤差より直感的だと思いますが、その場合、同じ結果になるのでしょうか?
  - なりません。詳しくは黒板で

質問・コメント（最急降下法）

- 最急降下法で局所解を求めることが出来るが、最適解は求まらないと考えました。最適解が求められるようにアルゴリズムに加える工夫が何かありますか?
  - 関数が凸関数の場合、求まります。詳しくは今回授業で
- 目標とするtが多項式で近似されるような場合、線形モデルでは適合しないが、このような場合、どのような手法をもって回帰を実現するのか?
  - それを多項式回帰といいます。詳しくは今回授業で。

そのほか

- 板書の字をもう少し大きくしてほしい
  - 大きくします
- 部屋が狭くて帰っている人が結構います
  - 部屋を大きくしました
- 線形代数と微分を復習しないといけないと感じた
  - はい、ぜひ復習してください

文書を特徴ベクトルに：Bag-of-words

文書

In recent years, there has been a growing trend towards outsourcing of computational tasks with the development of cloud services. We propose two building blocks that work with FHE: a novel batch greater-than primitive, and matrix primitive for encrypted matrices

辞書(1.2万語)

ID	word
1168	batch
1169	bath
⋮	⋮
1201	cloud
⋮	⋮
1239	computation
1240	computational
⋮	⋮
1172	primitive

頻度は考えずに出現したかどうかを特徴とする

$$\boldsymbol{x}_i = (0, 0, \ldots, 0, 1, 0, 1, 0, \ldots, 0, 1, 0, \ldots, 0, 1, 0, \ldots)$$

Using Fully Homomorphic Encryption for Statistical Analysis of Categorical, Ordinal and Numerical Data
Wenjie Lu, Shohei Kawasaki, Jun Sakuma, IACR Cryptology ePrint Archive, Informal and Other Publications (2016)

文書を特徴ベクトルに:

日本語の場合は形態素解析が必要

文書における特徴ベクトルの要素は文書中の単語です

- 文書 名詞,一般,*,*,*,*,文書,ブンショ,ブンショ
- における 助詞,格助詞,連語,*,*,*,における,ニオケル
- 特徴 名詞,一般,*,*,*,*,特徴,トクチョウ,トクチョー
- ベクトル 名詞,固有名詞,一般,*,*,*,*,
- の 助詞,連体化,*,*,*,*,の,ノ,ノ
- 要素 名詞,一般,*,*,*,*,要素,ヨウソ,ヨーソ
- は 助詞,係助詞,*,*,*,*,は,ハ,ワ
- 文書 名詞,一般,*,*,*,*,文書,ブンショ,ブンショ
- 中 名詞,接尾,副詞可能,*,*,*,中,チュウ,チュー
- の 助詞,連体化,*,*,*,*,の,ノ,ノ
- 単語 名詞,一般,*,*,*,*,単語,タンゴ,タンゴ
- です 助動詞,*,*,*,特殊・デス,基本形,です,デス,デス

形態素解析

頻度表

単語	出現頻度
文書	2
における	1
特徴	1
ベクトル	1
の	2
要素	1
は	1
中	1
単語	1
です	1

Stop word除去

単語	頻度
文書	2
特徴	1
ベクトル	1
要素	1
単語	1

Bag-of-words

http://www.nal.ie.u-ryukyu.ac.jp/wiki/index.php?yono%2FPython%A4%C7%BC%AB%C1%B3%B8%C0%B8%EC%BD%E8%CD%FD

TF-IDF

- Bag-of-wordsベクトルの”0”,”1”の代わりに語の出現頻度や語の珍しさを反映した値を入れる
- N個の文書 $D = \{d_1, d_2, \ldots, d_N\}$
  - n_{ij} : j番目の文書における語iの出現回数
- TF (term freq.)

$$\mathrm{tf}_{ij} = \frac{n_{ij}}{\sum_{d_k \in D} n_{kj}}$$

語iの文書jにおける出現割合

- IDF(inv. doc. freq.)

$$\mathrm{idf}_i = \frac{N}{|\{d|^i \ \in d, d \in D\}|}$$

語iの珍しさ

- TF/IDF

$$\mathrm{tfidf}_{ij} = \mathrm{tf}_{ij} \times \mathrm{idf}_i$$

語iの文書jにおける重要さ

画像を特徴ベクトルに

- 認識対象について，必要な情報を残し不要な情報を削除する
- 顔認識⇒ **輝度値特徴**
  - 有効：顔の見え方そのもの
  - 不要：画像の明るさ（照明によって明るさは変化）

- 物体認識⇒ **高次局所自己相関特徴**
  - 有効：見え方，色，形状，個数
  - 不要：物体の位置

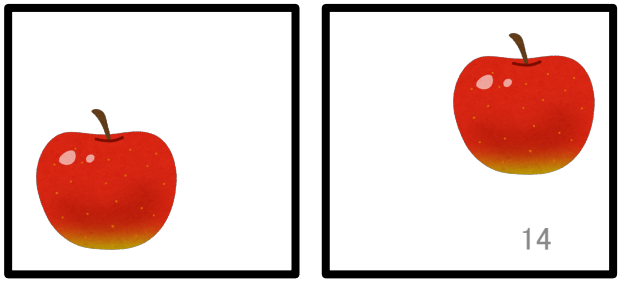


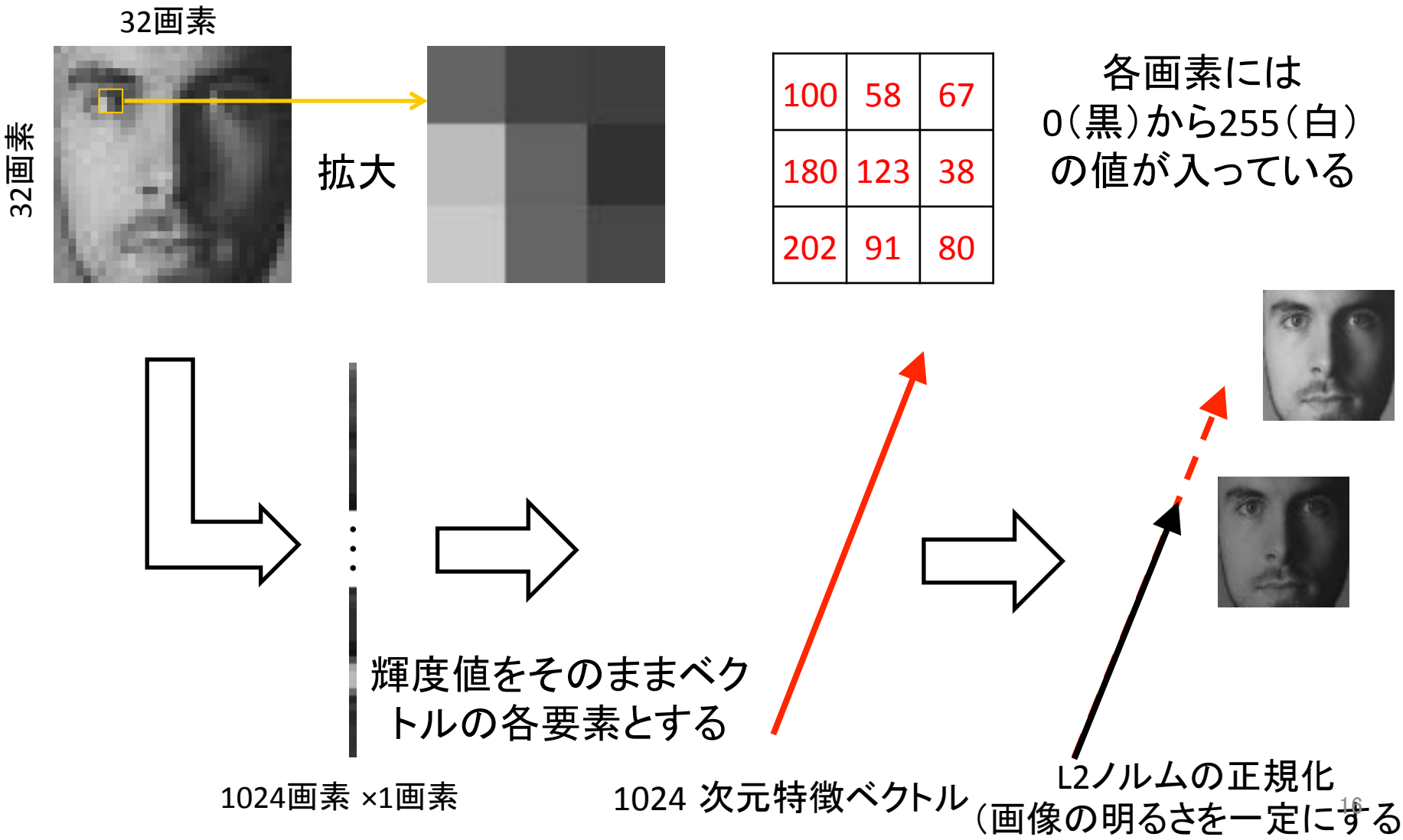
Photo by Simon Robben from Pexels

画像を特徴ベクトルに

- 輝度値特徴
  - アピアランスベース画像認識と呼ばれる
  - 画像の見え方をそのまま用いる方法
  - 画素×画素次元特徴ベクトルになる
- 高次局所自己相関特徴[1]
  - 位置不変性，加法性，低次元，高速，ロバスト等の画像認識に適した特性を持つ
  - 様々な応用（画像の印象評価，ジェスチャ認識等），拡張（カラー画像，動画像等）が提案されている

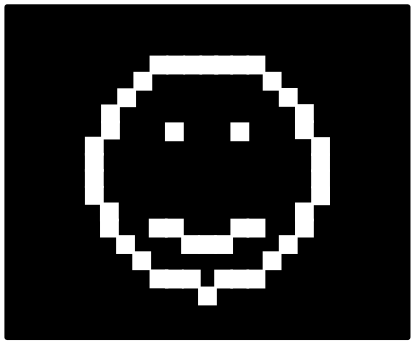
[1] 小林匠, 大津 展之, “画像特徴量 – 高次局所自己相関に着目した画像特徴量と画像認識への応用”, 電子情報通信学会誌, 2011

輝度値特徴

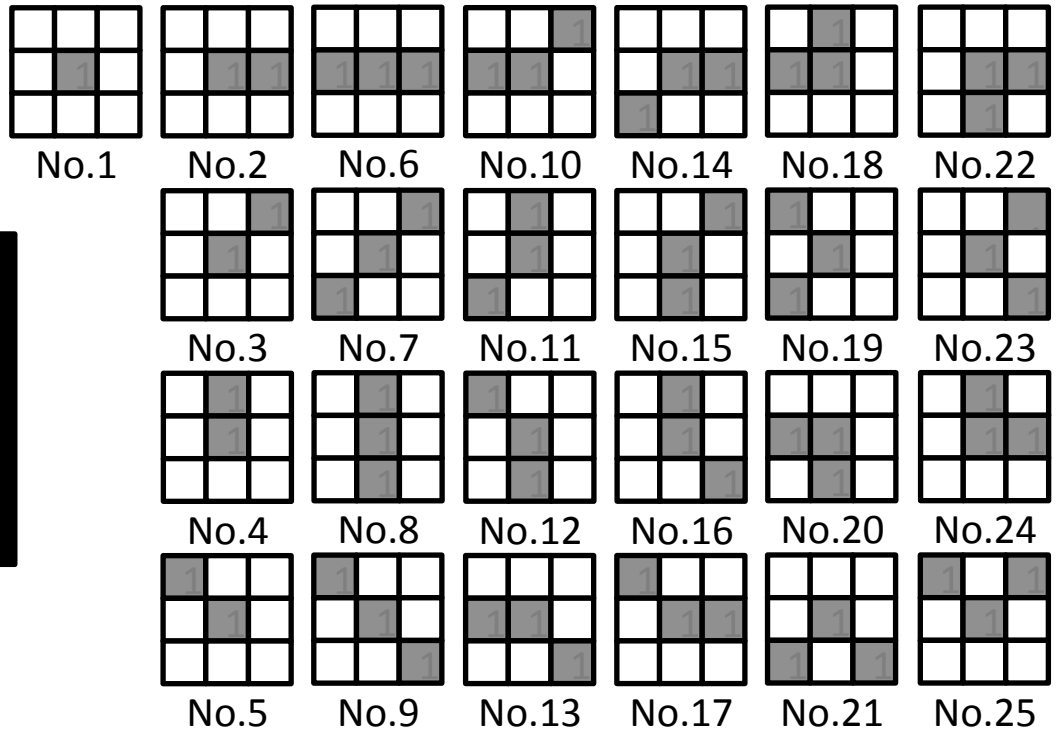

32画素
32画素
拡大
100 58 67
180 123 38
202 91 80
各画素には
0（黒）から255（白）
の値が入っている
輝度値をそのままベクトルの各要素とする
1024画素 ×1画素
1024 次元特徴ベクトル
L2ノルムの正規化
（画像の明るさを一定にする）

高次局所自己相関特徴 (HLAC)

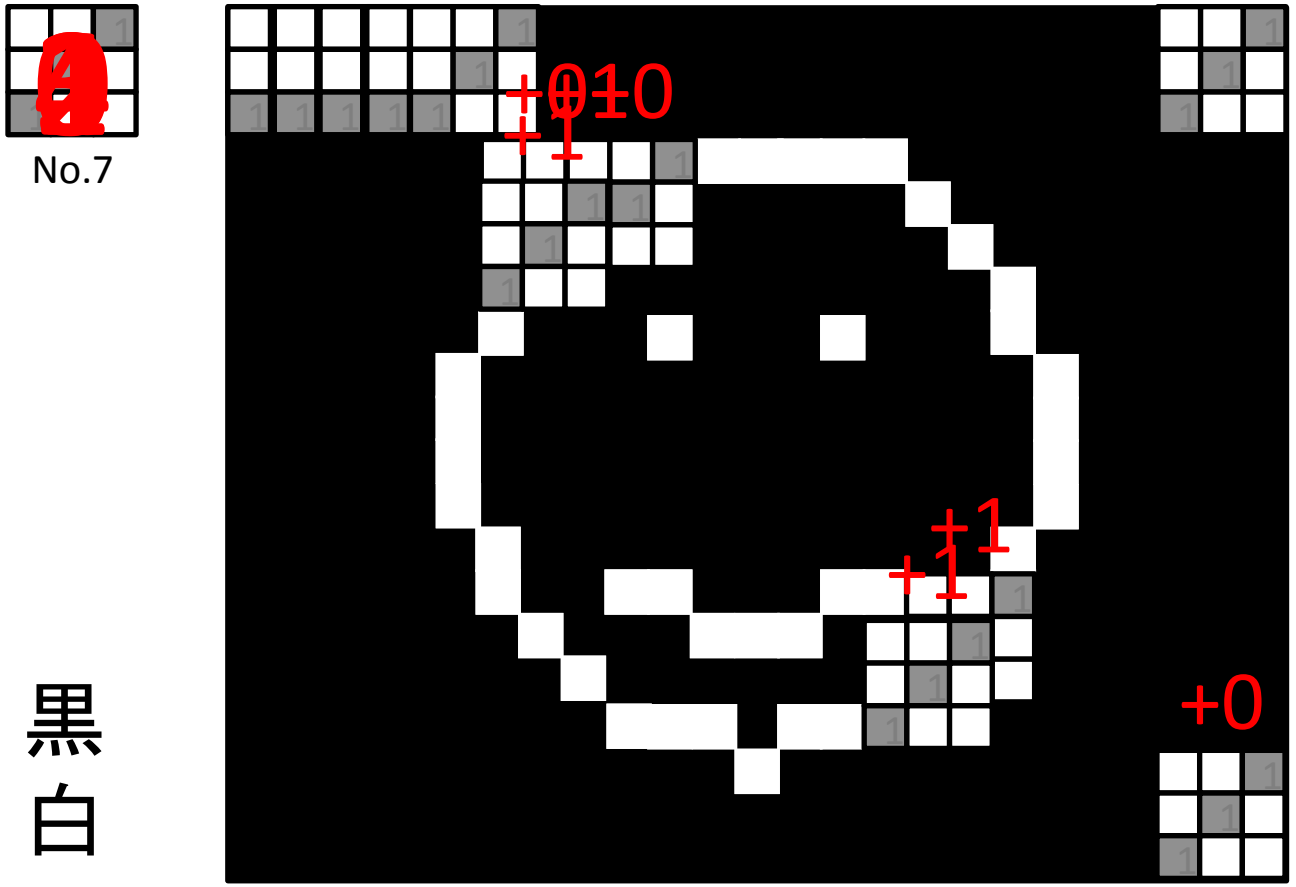
- 各画素において，マスクの灰色の箇所と重なる画素値の積を計算，画面全体に走査しその和を計算
- 画像のサイズに依らず25次元特徴ベクトルとなる



入力画像



HLAC特徴の計算法（2値画像の場合）

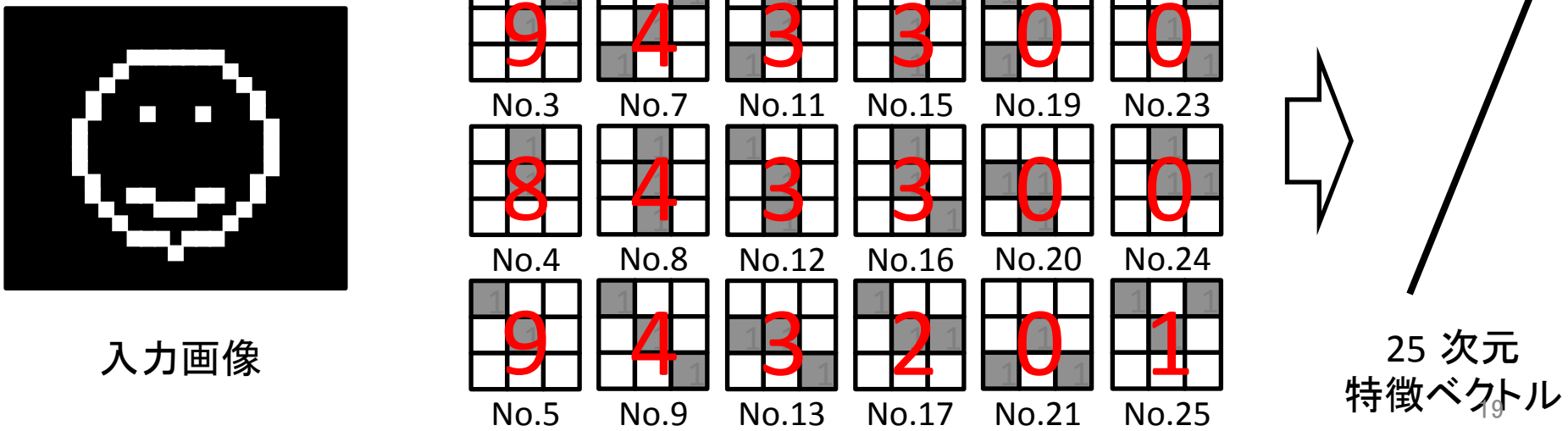


0：　黒

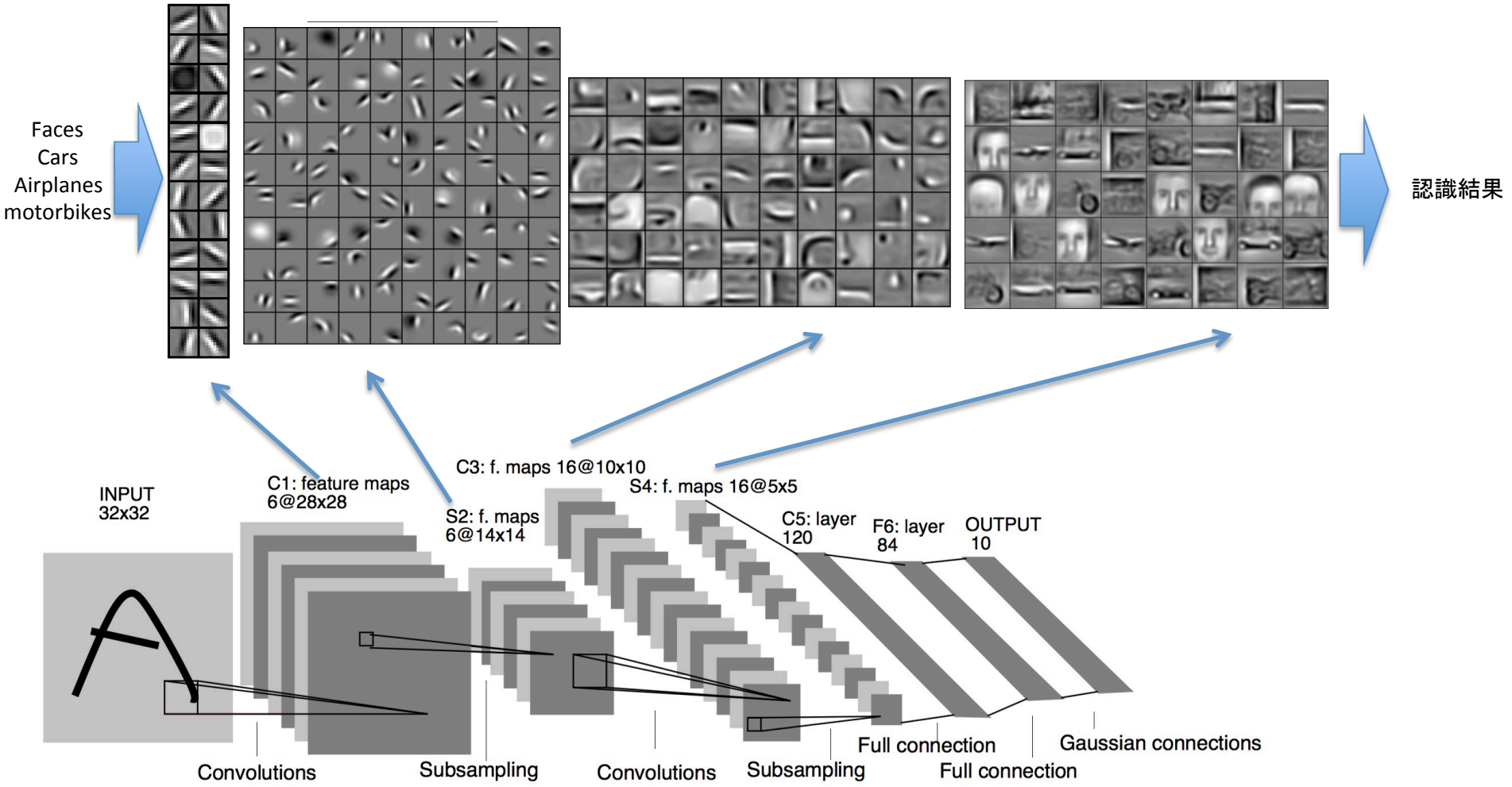
1：　白

HLAC特徴の計算法（2値画像の場合）

- 2値画像の場合，各マスク（直線，曲線など）が画像中に何個あるかを数え上げることに相当する



Deep learningによる特徴発見



Lee et al. “Convolutional DBNs for Scalable Unsupervised Learning of Hierarchical Representations” ICML09

特徴量Q&A

- Q. 画像パターンをマッチングさせる特徴量で，必要なパターンが用意できなかったらどうするか？
  - A. どんなデータも表現できる特徴量というのは実現困難なので，目標を定めてある程度まではがんばるがそれ以上はあきらめる，ということになります
- Q. どんなデータもうまいことベクトルに変換できる？
  - A. いい変換を見つけるのは結構職人技です
  - A. この職人技をも機械にやらせるのがdeep learningです

機械学習の流れ

データ

新元号「令和」のもとで初めてとなる予算編成に向けて国の財政制度等審議会は、予算規模の拡大が続く中、社会保障費などの歳出の抑制策を検討する新たな部会を設け、集中的に議論を行うことになりました。
財務大臣に予算の在り方などを提言する財政制度等審議会は、4日から「令和」のもとで初めてとなる、来年度の予算編成に向けて議論を始めました。

深層学習

特徴量(データのベクトル表現)

$$X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\},$$
$$\boldsymbol{x}_i \in \mathbb{R}^D$$

(浅い)機械学習

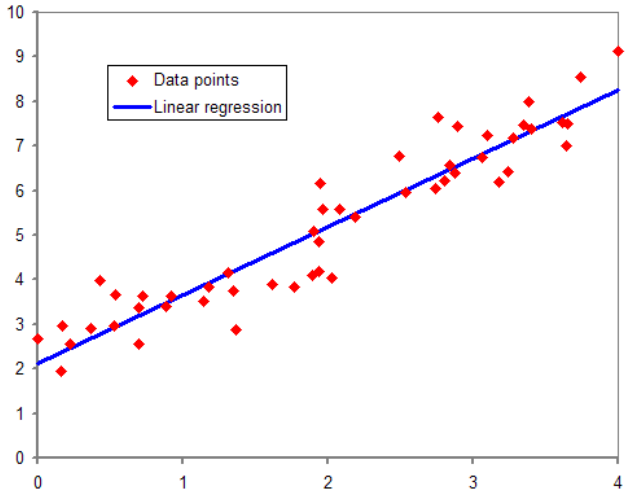
機械学習(教師つき, 教師なし)
特徴量から概念への写像

概念の獲得
スパム、ネコ, 政治ニュース, etc.

https://unsplash.com/photos/IbPxGLgJiMI
Erik-Jan Leusink

NHK NEWS WEB

線形回帰(重回帰)



- 特徴(独立変数) $\boldsymbol{x}_i \in \mathbb{R}^D$
- 目標値(従属変数) $t_i \in \mathbb{R}^1$
- 事例:特徴と目標値の組 $(\boldsymbol{x}_i, t_i), i = 1, \ldots, N$
- 目的:多数の目標値(教師)付き事例が与えられたときに、事例から目標値を予測する

https://commons.wikimedia.org/wiki/File:Normdist_regression.png

データの表現

- 事例の行列表現

慣例として、一つのサンプルを、
行列の行で表す

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_N^T \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{ND} \end{pmatrix}$$

- 目標値のベクトル表現

$$\boldsymbol{t} = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{pmatrix}$$

線形回帰の定式化

- 線形回帰モデル
  - 単回帰　$t = w_0 + w_1 x$
  - 重回帰　$t = w_0 + w_1 x_1 + \ldots + w_D x_D = w_0 + \sum_{d=1}^{D} w_d x_d$
    - 複数の特徴量とバイアスから目標値を予測

- 事例を

先頭に1を付け加える

$$\begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_N^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1D} \\ 1 & x_{21} & x_{22} & \cdots & x_{2D} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{ND} \end{pmatrix}$$

と定義すると

$$t = \sum_{i=0}^{D} w_i x_i = \boldsymbol{w}^T \boldsymbol{x}$$

重回帰モデルが
内積一発で書けて便利

wine dataの線形回帰

特徴量x　　目標値t

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

- 単回帰 (特徴量を一つだけ使う)
  - quality = w_0+ w_1* fixed_acidity
- 重回帰(複数の特徴量をつけう)
  - quality = w_0+w_1* fixed_acidity+ w_2*volatile_acidity + … + w_{11}*alcohol

wine dataの線形回帰

- [code] wine_linearRegressionManyFeatures.ipynb

負の影響

正の影響

	Coefficients	Name
1	-0.193967	volatile acidity
6	-0.107356	total sulfur dioxide
4	-0.088183	chlorides
8	-0.063842	pH
2	-0.035553	citric acid
7	-0.033737	density
3	0.023019	residual sugar
0	0.043497	fixed acidity
5	0.045606	free sulfur dioxide
9	0.155277	sulphates
10	0.294243	alcohol

Mean squared err. 0.41676716722140794

MSEは最良の単回帰より改善

二乗誤差最小化

- 二乗誤差の和 $E(\boldsymbol{w}) = \sum_{i=1}^{N} (t_i - \boldsymbol{w}^T \boldsymbol{x}_i)^2$
- 二乗誤差の和を最小にするモデルパラメータ

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \sum_{i=1}^{N} (t_i - \boldsymbol{w}^T \boldsymbol{x}_i)^2$$

  - 1/Nはなくても結果は変わらないことに注意
  - $\arg\min_x f(x)$ 関数fを最小にする引数x
  - w*をどうやって求めるか？

最適化の定義

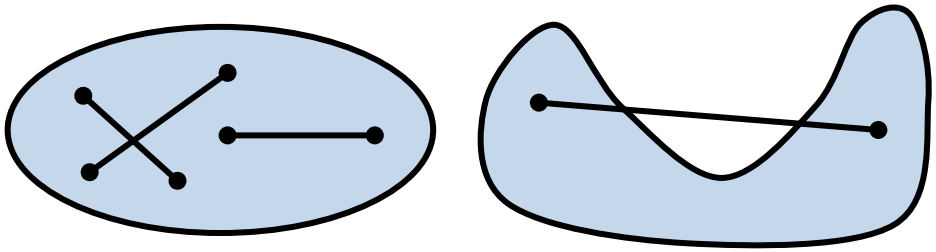
$$\text{maximize } f(x) \qquad \text{目的関数}$$

$$\text{subject to } g(x) = 0 \qquad \text{等式制約}$$

$$h(x) \leq 0 \qquad \text{不等式制約}$$

- 制約を満たす解：実行可能解
- 制約を満たす領域：実行可能領域

凸集合と凸関数

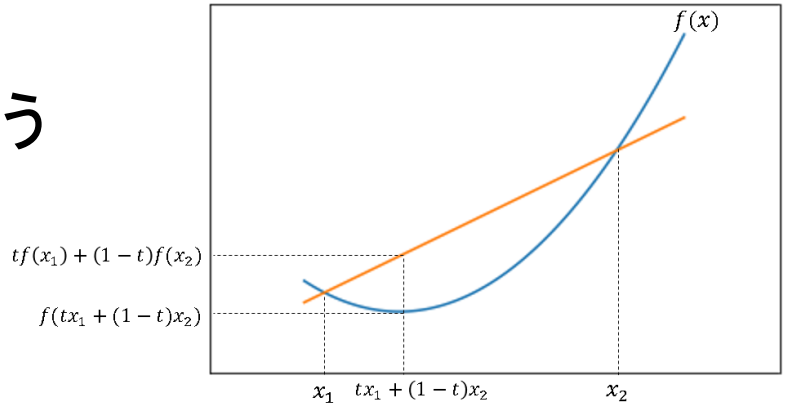


(a)凸集合　　(b)非凸集合

- 凸集合と非凸集合
- 集合Aが凸集合⇔

$$\forall x, y \in A \text{ and } t \in [0,1],$$
$$tx + (1-t)y \in A$$

- 下に凸な関数：凹関数ともいう



- 関数fが下に凸

⇔ $\forall t \in [0,1], f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$

どんな関数が凸関数か

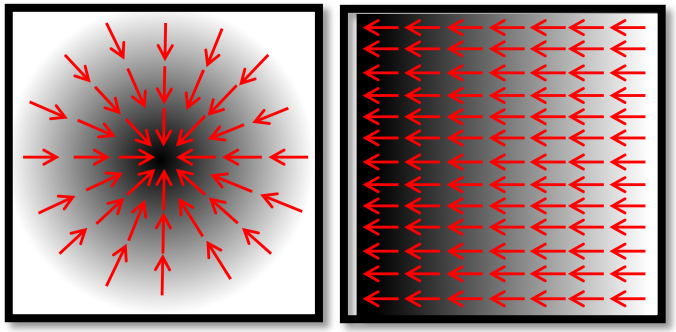
- 線形関数，アフィン関数
- 正定値行列Q について $x^T Q x + c^T x$
- $exp(x), -\log(x), x\log(x)$
- ノルム(あとでやります)
- $x_i \geq 0$ について $(x_1 x_2 \ldots x_n)^{1/n}$
- $\log(e^{x_1} + e^{x_2} + \ldots + e^{x_n})$

再掲

行列代数基礎:多変数関数の微分

- 関数　$f : \mathbb{R}^D \to \mathbb{R} \quad f(\boldsymbol{x}) = a$
- ある変数(x_d)での微分　$\dfrac{\partial f}{\partial x_d}$
- すべての変数での微分(勾配)

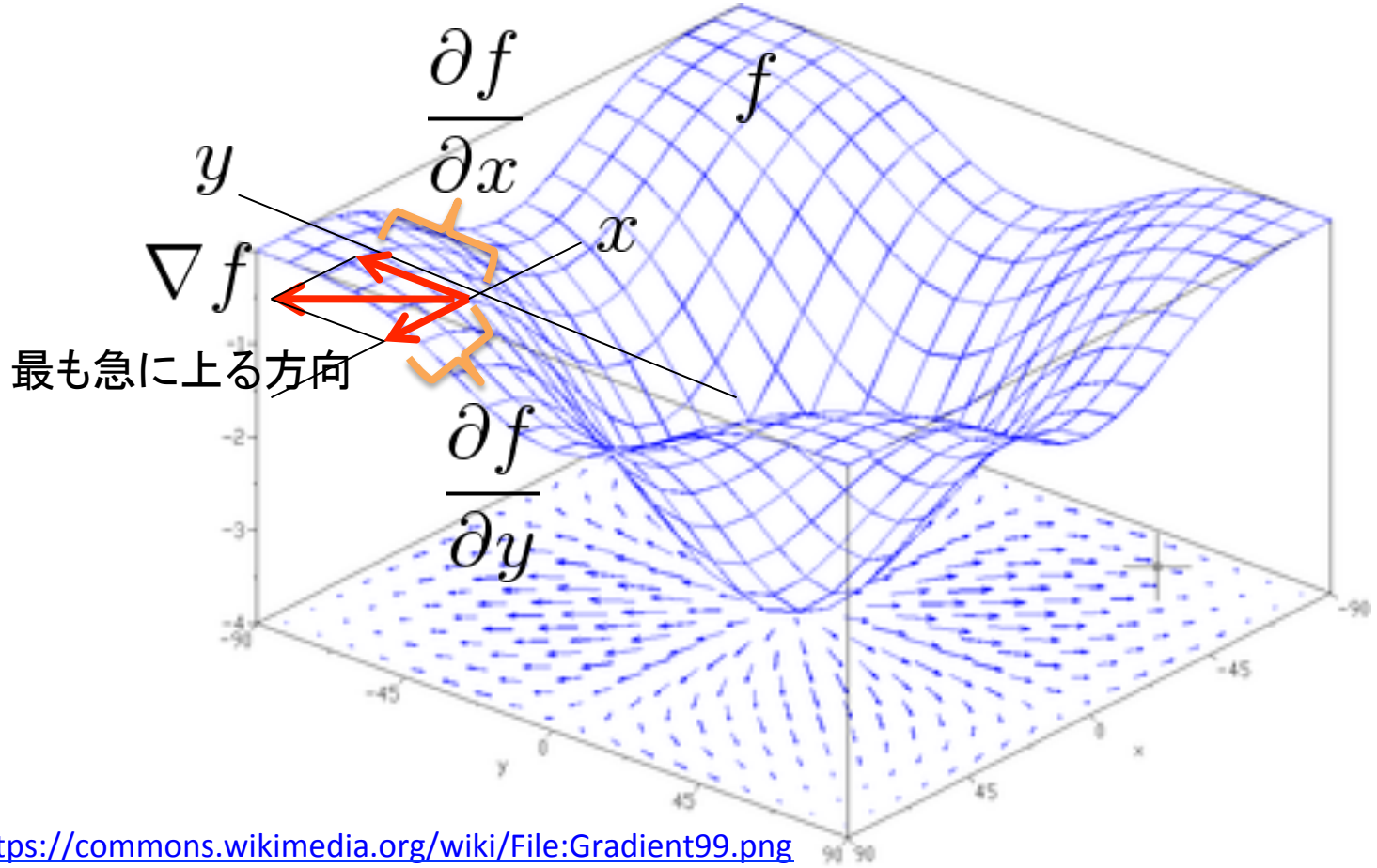
$$\frac{\partial f}{\partial \boldsymbol{x}} = \nabla_{\boldsymbol{x}} f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{pmatrix}$$



勾配

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_d} \right)^T$$

- $f(x,y) = -(\cos^2 x + \cos^2 y)^2$ の勾配を底面に射影



https://commons.wikimedia.org/wiki/File:Gradient99.png

参考

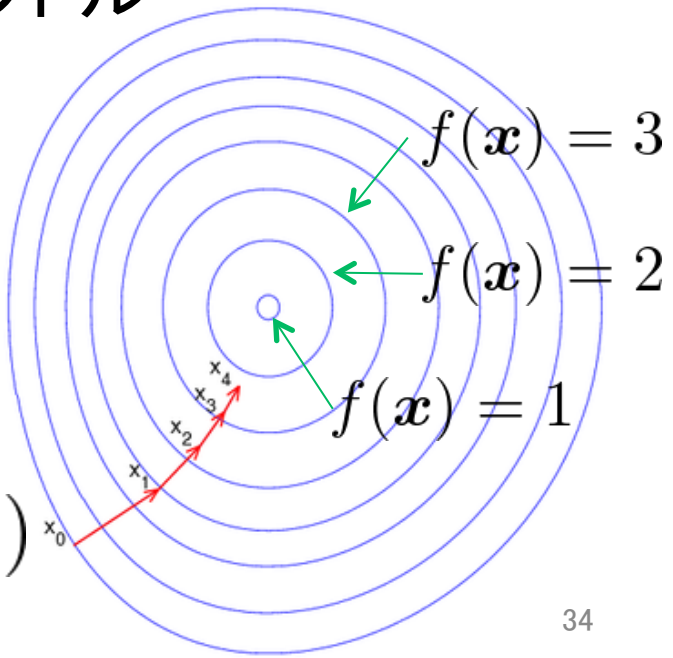
等高線と勾配

- 等高線：同じ $f(\boldsymbol{x})$ の値を与える $\boldsymbol{x}$ の集合
  - 閉曲線になる
- 勾配：点 $\boldsymbol{x}$ において最も大きく $f(\boldsymbol{x})$ が増える方向
  - D次元関数の勾配はD次元ベクトル

勾配 $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_d} \right)^T$

勾配のサイズ＝坂の急さ $\left\| \nabla_{\boldsymbol{x}} f(\boldsymbol{x}) \right\|$

値が減っているので，
勾配の逆方向 $-\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0)$



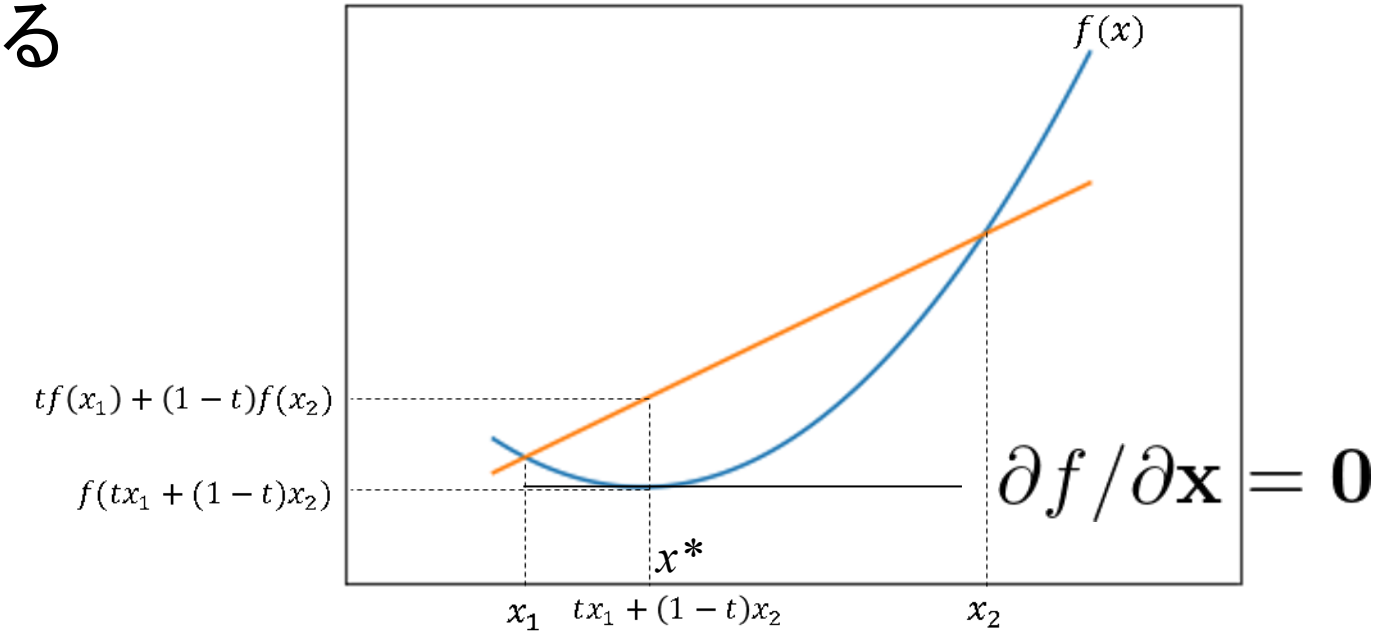
https://commons.wikimedia.org/wiki/File:Gradient_descent.svg

微分可能な凸関数の最適化

1. 解析的な解法
    1. 勾配ベクトルが0ベクトルとなるxを解析的に求める
    2. 解析的な解法が使える場合は、厳密な解が求まるため、可能ならばこちらを使う
2. 近似的な解法
    1. 最急降下法 (とその仲間)
    2. ランダムに初期化し、勾配方向に解を繰り返し更新
    3. 近似解が求まる (使った時間に応じて精度が改善)
    4. 解析的な解法が使える場合でも、近似解法が利用されることがある (計算上の理由による、詳しくは後ほど)

解析的な解法
微分可能な凸関数の最小化

- 関数 $f : \mathbb{R}^D \to \mathbb{R}$ は微分可能かつ凸関数とする
- f はただ一つの極値を持ち、それを与える $\boldsymbol{x}$ は $\partial f / \partial \mathbf{x} = \mathbf{0}$ を満たす
- そのようなxをx*とおけば、fの最小値は$f(x$*$)$で与えられる



最急降下法 (近似解法)

- 問題 $\min_{\boldsymbol{w}} f(\boldsymbol{w})$ なるwを求める
  - 関数fは微分可能
- アルゴリズム
  1. $t = 0, \boldsymbol{w}^0$ をランダムに初期化
  2. $\boldsymbol{w}^{t+1} \leftarrow \boldsymbol{w}^t - \eta \nabla_{\boldsymbol{w}} f$
  3. $\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\| < \epsilon$ ならば停止、そうでなければstep 4へ
  4. $t \leftarrow t + 1$ としてstep2へ

ステップサイズパラメータη：1ステップにどの程度更新させるか
収束判定パラメータ ϵ

解析演習：凸関数の最適化

$f(\boldsymbol{x}) = 2x_1^2 + x_1x_2 + x_2^2 - 5x_1 - 3x_2 + 4$ とする。$f(\boldsymbol{x})$ が凸関数であることは既知とする。

1. f の勾配 ∇f を求めよ
2. $(0,0), (1,2), (1,0.5), (1,1)$ における f の勾配を求めよ
3. f を最小にする $\boldsymbol{x}$ とその時の $f(\boldsymbol{x})$ を求めよ

凸計画問題

- 凸計画問題
  - 目的関数が凸関数
  - 実行可能領域が凸集合

凸計画問題	制約なし	等式制約	不等式制約
fが微分可能	fの微分を0にする点が解析的に求まる ⇒その点が最適解 求まらない ⇒最急降下法など	ラグランジュの未定乗数法 6,7週後に触れます	ラグランジュの未定乗数法 今回は扱わない (SVM導出などで出てくる)
fが微分不可能	2,3週後に触れます	今回は扱わない	今回は扱わない

凸関数でない関数の最適化は、9-10週あたりで扱います。

凸関数Q&A

- xlogxの概形を示してほしい
  - gnuplot http://gnuplot.respawned.com/ を使えば以下のコマンドで出ますp x*log(x)
- 二階微分で上に凸か下に凸か判別できると学んだが？
  - 1変数関数の場合はそれで判定できますが，2変数以上の関数の場合，それだけでは判別できません．のちに出てくる、ヘシアン行列を使って判定します。具体的には、全ての点においてヘシアン行列が正定値行列(全ての固有値が正)であれば、(狭義)凸関数だといえます

最適化Q&A

- 凸計画問題で「微分可能かつ制約なし」で最適解が求まらないものにはどんなものが?
  - 微分はできたけど，それを0にする解が解析に解けない場合，になります．そのような場合は，最急降下法をつかいます．ロジスティック回帰でててきます
- 凸関数でない場合の最適化では，勾配方向に下っていけばよい？その際に局所解に陥らないためには？
  - そのとおりです．局所解に陥らないための工夫はいろいろありますが、本質的な解決策はありません。ニューラルネットワークの学習は、凸関数でない場合の最適化になりますが、経験的に(確率的)最急降下法で比較的うまくいくことがわかっています。その理由はまだあまりよくわかっていません。

二乗誤差和の最小化

- 二乗誤差和 $$E(\boldsymbol{w}) = \sum_{i=1}^{N} (t_i - \boldsymbol{w}^T \boldsymbol{x}_i)^2$$
- 二乗誤差の最小化
  - $E(\boldsymbol{w})$は$\boldsymbol{w}$ に関して下に凸な関数
    - 下に凸であることは前提とする*
  - $\dfrac{\partial E}{\partial \boldsymbol{w}} = 0$ になるような$\boldsymbol{w}$を求めればよい
  - $\boldsymbol{w} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{t}$ (導出は演習)

*Eのヘシアン行列の固有値が全て正であることを示せば良い

解析演習：線形回帰の導出

$\boldsymbol{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{iD} \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_N^T \end{pmatrix}, \boldsymbol{t} = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{pmatrix}, \boldsymbol{w} = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{pmatrix}$ とする。事例群 $\{(\boldsymbol{x}_i, t_i)\}_{i=1}^N$ を使って線形回帰モデル $t = \boldsymbol{w}^T \boldsymbol{x}$ を求めることを考える。

1. 二乗誤差和 E を $\boldsymbol{w}$ の関数で表せ
2. 二乗誤差和 E の $\boldsymbol{w}$ についての勾配 $\nabla_{\boldsymbol{w}} E$ を求めるために、以下を導出せよ
   (a) $\sum_{i=1}^N t_i \boldsymbol{x}_i = \boldsymbol{X}^T \boldsymbol{t}$
   (b) $\sum_{i=1}^N \boldsymbol{x}_i \boldsymbol{w}^T \boldsymbol{x}_i = \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{w}$
3. 勾配 $\frac{\partial E}{\partial \boldsymbol{w}}$ を $\boldsymbol{x}_i$(あるいは $\boldsymbol{X}$), $\boldsymbol{t}$ の式で示せ。
4. (近似解法) 線形回帰モデルを最急降下法で求めるときの、パラメータの更新式を $\boldsymbol{x}_i$(あるいは $\boldsymbol{X}$), $\boldsymbol{t}$ の式で示せ。初期解を $\boldsymbol{w}^0$, t 回目の更新時の回を $\boldsymbol{w}^t$, ステップサイズパラメータを η とする。
5. (解析解) $\frac{\partial E}{\partial \boldsymbol{w}} = 0$ なる $\boldsymbol{w}$ が $\boldsymbol{w} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{t}$ であることを示せ。

回帰ここまでのまとめ

- xからtを線形式で予測する $t = \sum_{i=0}^{D} w_i x_i = \boldsymbol{w}^T \boldsymbol{x}$
- 回帰係数wによる予測の良さは二乗誤差の小ささで評価 $E(\boldsymbol{w}) = \sum_{i=1}^{N} (t_i - \boldsymbol{w}^T \boldsymbol{x}_i)^2$
- 二乗誤差はwについて凸なので，二乗誤差を最小にするwはE(w)をwで微分してゼロになるところ

$$\frac{\partial E}{\partial \boldsymbol{w}} = 0 \Rightarrow \boldsymbol{w} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{t}$$

線形回帰Q&A

- Q. 回帰の結果が曲線になるのはどんな場合？
  - A. 線形回帰は全て直線になります．曲線への一般化は来週．直線は曲線を含むことに注意．
- Q. どのようなデータをもってくるかで，最終的な評価が変わってくる？
  - A. 鋭い．データのばらつきや，数で結果は変わってきます．
- Q. 縦軸はy?
  - A. 予測を表す目標値の変数をtにしています
- Q. 最終二乗誤差はどれくらいの値の範囲にあるのか？
  - A. 原理的には無限に大きくなります．
- Q. 最小二乗法と一緒か
  - A. 一緒です