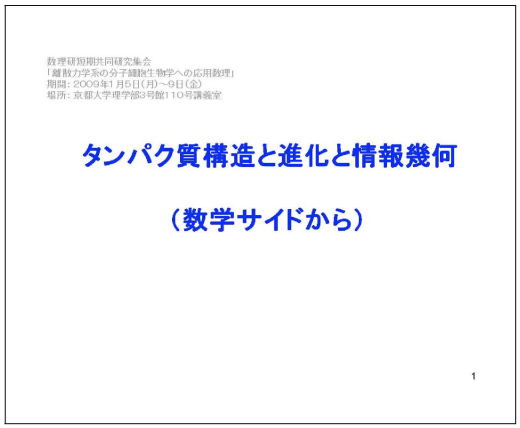
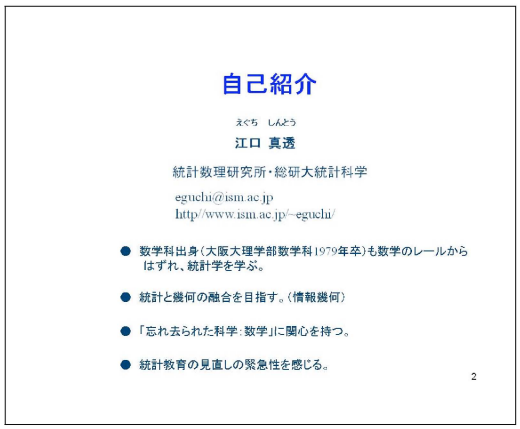
数理解析研究所講究録
第 1698 巻 2010 年 205-218

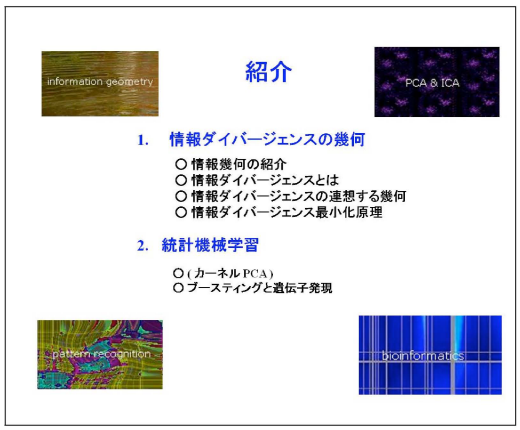


- The 1st MSJ-SI: Probabilistic Approach to Geometry (2008.7.28-8.8, 京都大学) での招待講演

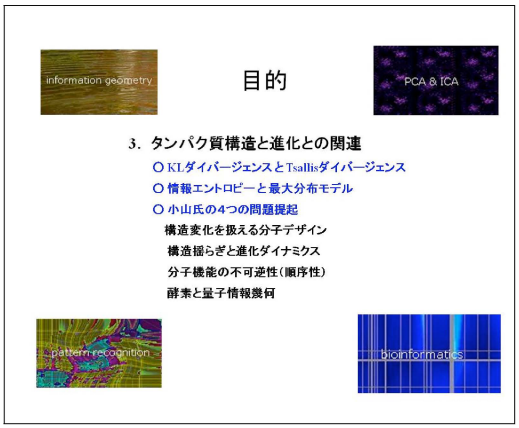
  "Information divergence geometry and its application to machine learning"
  が元になっている.



- 大学院から統計の勉強をする.
- 甘利俊一氏が情報幾何を始めた時期で，プレプリントを見て興味を持ち修士のテーマを変える.
- 社会的に見て数学は他分野に比べ進展していない.
- 統計学科は日本にない．したがって日本で統計学を学部で専門的に勉強することは難しい.
- 日本の国家戦略として統計学が抜けているのはマイナスではないか．少なくとも統計学の教育は大事にすべき.

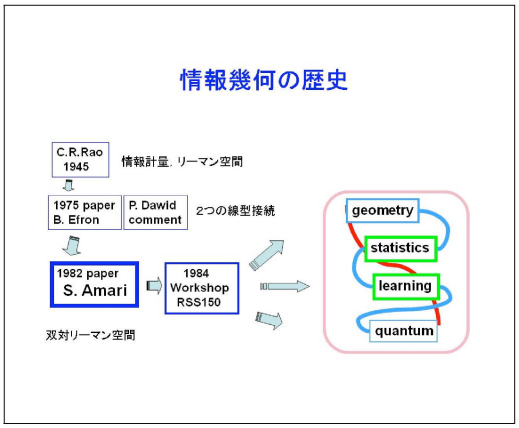


- カーネル PCA の話は時間の都合上省略.

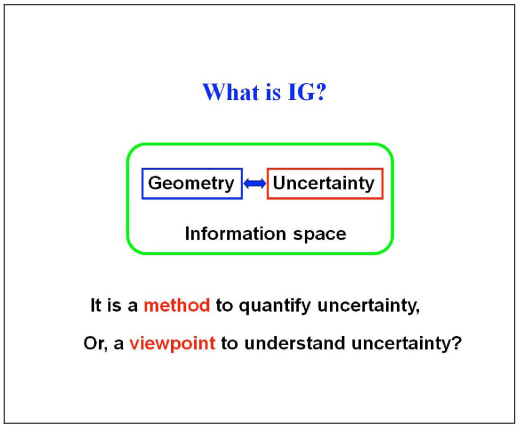


目的は数学と生物の融合

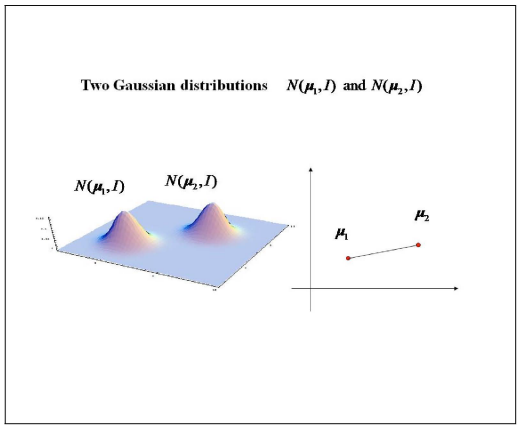
- 小山氏の講演にあった KL ダイバージェンスの最大化の問題を情報幾何の観点から見ていきたい．
- 小山氏による 4 つの問題は情報幾何の方法論のバラエティの中で扱えるのではないか．



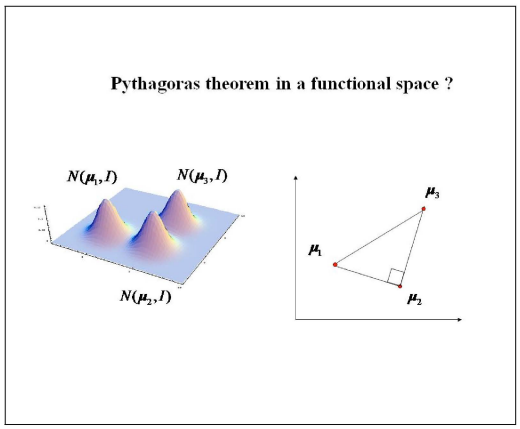
- 歴史的に見て統計学の中での最初の幾何的な考え方は Gauss の最小二乗法．線形射影という話から Gauss norm の定義が成立する．
- 微分幾何的な考え方は C.R. Rao が最初ではないか (Bull. Calcutta Math. Soc., 1945)．情報計量を元にして統計モデルはリーマン空間になる．
- 実はリーマン空間では不十分．２つの線形接続が必要になる (Efron, Dawid)．
- 甘利氏により標準的な微分幾何の言葉で定式化．



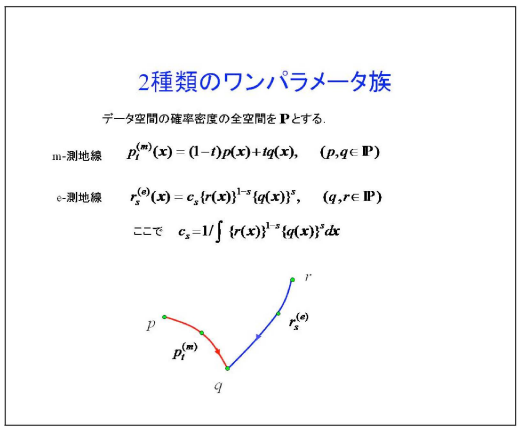
- 情報幾何は情報空間の中での不確定性を幾何的に sharp に理解することを目指している．



- 統計モデルは分布の集まりである．規格化することにより分布をユークリッド空間内の点と思うと，2 つの分布を結ぶ path は 2 点を結ぶ path に相当する．



- 3 つの分布があったとき，対応するユークリッド空間内の 3 点の関係，たとえばピタゴラスの定理が考えられる．



上記の一般化

- 3 点 p, q, r に対し，p と q を結ぶ測地線および r と q を結ぶ測地線を結び，ピタゴラスの定理を考えてみる．

自己紹介

えぐち　しんとう
江口 真透
統計数理研究所・総研大統計科学
eguchi@ism.ac.jp
http://www.ism.ac.jp/~eguchi/

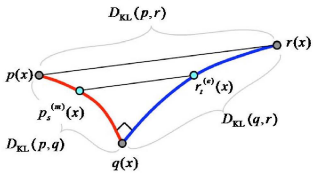
- 数学科出身（大阪大理学部数学科1979年卒）も数学のレールからはずれ、統計学を学ぶ。
- 統計と幾何の融合を目指す。（情報幾何）
- 「忘れ去られた科学：数学」に関心を持つ。
- 統計教育の見直しの緊急性を感じる。

2

Kullback-Leibler(KL) ダイバージェンス

- 正規分布に対し，KL ダイバージェンスはユークリッド距離の 2 乗に比例するのでピタゴラスの定理と同じ意味になる．
- すなわち，3 点を同一平面上にとるユークリッド幾何に帰着される．

Pythagorean Theorem



$$\dot{p}_t^{(m)} \perp_g \dot{r}_s^{(e)} \quad \Rightarrow \quad D_{KL}(p_t^{(m)}, r_s^{(e)}) = D_{KL}(p_t^{(m)}, q) + D_{KL}(q, r_s^{(e)})$$

$$D_{KL}(\vec{p}_t, \vec{r}_s) \geq D_{KL}(q, \vec{r}_s)$$

$$D_{KL}(\overleftarrow{p}_t, \vec{r}_s) \geq D_{KL}(\overleftarrow{p}_t, q) \quad \text{(em algorithm, Amari, 1995)}$$

13

- 2 種類の測地線に対し，ピタゴラスの定理が成立するための必要十分条件は 2 つの path の点 q での接ベクトルが直交していることである．
- これにより 2 つの不等式，いわゆる em アルゴリズムが簡単に導かれる．

Information Geometry

Statistical model $M = \{q(x,\theta) : \theta \in \Theta\}$

where $q(x,\theta)$ is a pdf st $\int q(x,\theta)\,dx = 1$

Information metric (Fisher, 1922; Rao, 1945)

$$g_{ij}(\theta) = E_{q(x,\theta)}\{e_i(x,\theta)e_j(x,\theta)\} \quad \text{where} \quad e_i(x,\theta) = \frac{\partial}{\partial\theta^i}\log q(x,\theta)$$

Dual connections (Amari, 1982)

Exponential connection $\overset{e}{\Gamma}_{ij,k}(\theta) = E_{q(x,\theta)}\left(\frac{\partial^2}{\partial\theta^i\partial\theta^j}\log q(x,\theta)e_k(x,\theta)\right)$

Mixture connection $\overset{m}{\Gamma}_{ij,k}(\theta) = \int \frac{\partial^2}{\partial\theta^i\partial\theta^j} q(x,\theta)e_k(x,\theta)dx$

$\frac{1}{2}(\overset{e}{\Gamma}_{ij,k}(\theta)+\overset{m}{\Gamma}_{ij,k}(\theta))$ is Riemannian connection with $\{g_{ij}(\theta)\}$

9

- 統計モデル M を • でパラメトライズされた分布の集まりとする．
- 情報計量について
  - Fisher がフィッシャー情報量を定義．
  - Rao が，これを計量とみなすことによりリーマン空間となることを指摘．
  - この量はハイゼンベルグの不確定性原理とほとんど等価で推定の限界を与える．
- 2 つの接続 (e-接続，m-接続) が古典的な係数で定義される．
  - 統計モデルの対数密度の二階微分と接ベクトルの内積
  - 密度の二階微分
- 情報幾何はこの 2 つの接続の組で一つの測地性を与える．
- この 2 つの接続の平均をとることにより g のもとでのリーマン接続になる．

Mixture and exponential models

Mixture model (mixture geodesic space)

$$M=\{q(x,\theta):\theta\in\Theta\}$$

$$q(x,\theta)=\sum_{i=0}^{K}\theta_i q_i(x)\ ,\quad \Theta=\{(\theta_0,\dots,\theta_K):\theta_i>0,\sum_{i=0}^{K}\theta_i=1\}$$

Cf. Maximum likelihood

Exponential model (exponential geodesic space)

$$M=\{q(x,\theta):\theta\in\Theta\}$$

$$q(x,\theta)=\exp\{\sum_{i=1}^{K}\theta_i t_i(x)-\psi(\theta)\},\quad \Theta=\{\theta:\psi(\theta)<\infty\}$$

where $\psi(\theta)=\log\int\exp\{\theta_i t_i(x)\}dx$

Cf. Gaussian, Bernoulli, Poisson, Gamma, …

10

2 つの接続に関する測地線・測地面

- どちらも分布 • は 2 種類の接続に対するアフィンパラメータになっている．
- 統計学とは統計モデルの上で統計的推定と統計的検定をする．モデルを作るということと，それで推論する，ということは別．したがって 2 種類の接続が必要．
  - モデル：対象
  - 推論：行為
- 対象と行為が組になって一つの統計という作業ができる．よって 2 種類の測地性が必要．
- これはアインシュタインの相対論 (重力場の幾何学) と大きく異なる点．

Information divergence geometry

$D: M\times M\to \mathbf{R}$ is an **information divergence** on a statistical model M

$\Longleftrightarrow$ (i) $D(p,q)\geqq 0$ with equality if and only if $p=q$

(ii) D is differentiable on $M\times M$

Let $\quad D(Z\cdots|XY\cdots)(p)=Z_p\cdots X_q Y_q\cdots D(p,q)|_{q=p}$

Then we get a **Riemannian metric** and **dual connections** on M (Eguchi, 1983,1992)

$$g^D,\quad g^D(X,Y)=-D(X|Y)$$

$$\nabla^D,\quad g^D(\nabla^D_X Y,Z)=-D(XY|Z)\quad(\forall Z\in\mathcal{X}(M))$$

$$^*\nabla^D,\quad g^D(^*\nabla^D_X Y,Z)=-D(Z|XY)\quad(\forall Z\in\mathcal{X}(M))$$

14

ダイバージェンスの定義

- KL ダイバージェンス以外のダイバージェンスも統一的に見ることができないか．
- 特に対角部分

  • = {(p, p) : p • M}

  で可微分であることが必要．
- ダイバージェンスはリーマン計量 g, 2 つの接続 • , ˚• を定める
- 微分 D(Z · · · |XY · · ·)(p) は Z と XY について非対称なので，• と ˚• は一般に異なる．
- KL ダイバージェンスの場合は
  - • ：e-接続
  - ˚• ：m-接続

  となり，2 種類の接続が KL の非対称性から出てくる．

Proof 1

g^D is a Riemann metric

(i) $g^D(X,Y)=g^D(Y,X)$

$D(X|\cdot)=0$ since $D(p,q)\geq D(p,p)=0$

$g^D(X,Y)-g^D(Y,X)=D(XY|\cdot)-D(YX|\cdot)=D([X,Y]|\cdot)=0$

(ii) g^D is positive - definite

$g^D(X,Y)=-D(X|Y)=D(XY|\cdot)$

because $D(Y|\cdot)=0$.

Proof 2

∇^D and $^*\nabla^D$ are dual connections

(i) $\nabla^D_{fX}Y=f\nabla^D_X Y$

(ii) $\nabla^D_X fY=f\nabla^D_X Y+(Xf)Y$

$$g^D(\nabla^D_{fX}Y-f\nabla^D_X Y,Z)=-D((fX)Y-fXY|Z)=0\quad(\forall Z)$$

$$g^D(\nabla^D_X(fY)-f\nabla^D_X Y-(Xf)Y,Z)=-D(X(fY)-(Xf)Y-fXY|Z)=0\quad(\forall Z)$$

(iii) Let $\bar\nabla=\frac{1}{2}(\nabla^D+{}^*\nabla^D)$.

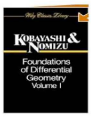
Then $Xg^D(Y,Z)=g^D(\bar\nabla_X Y,Z)+g^D(Y,\bar\nabla_X Z)$

$$Xg^D(Y,Z)-g^D(\bar\nabla_X Y,Z)-g^D(Y,\bar\nabla_X Z)$$
$$=-XD(Y|Z)+\frac{1}{2}\{XD(Y|Z)+XD(Z|Y)\}=0$$

16

参考書（微分幾何）

リーマン幾何学（近代数学講座）（単行本）
立花 俊一（著）単行本:
出版社: 朝倉書店; 復刊版 (2004/04)
ISBN-10: 4254116586



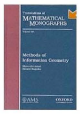
現代微分幾何入門（単行本）
野水 克己（著）
単行本:
出版社: 裳華房 (1981/01)
ISBN-10: 4785311274

Foundations of Differential Geometry (Wiley Classics Library) (ペーパーバック)
Shoshichi Kobayashi (著), Katsumi Nomizu (著)

どんな参考書（情報幾何）？

Differential Geometrical Methods in Statistics
Shun-Ichi Amari 1985 年 Springer

（http://www.geocities.jp/ismtstatseguchi/pdf/kokoro.pdf）



Methods of Information Geometry
Shun-Ichi Amari , Hiroshi Nagaoka
Amer Mathematical Society (2001)

U divergence

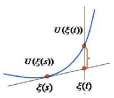
Let us take a convex function $U:\mathbb{R}\to\mathbb{R}$.

Let $u(z)=\frac{d}{dz}U(z)$, $u(\xi(z))=z$.

$$\xi(s)=\arg\max_{-\infty<t<\infty}\{ts-U(t)\}$$

$$u(t)=\arg\max_{-\infty<s<\infty}\{st-U^*(s)\}$$

where $U^*(s)=s\xi(s)-U(\xi(s))$



U divergence

$$D_U(p,q)=\int\{U(\xi(q))-U(\xi(p))-u(\xi(p))(\xi(q)-\xi(p))\}$$

$$=\int\{U(\xi(q))-p\xi(q)+U^*(\xi(p))\}$$

19

U-ダイバージェンスの構成法

- • は u の逆関数.
- U-ダイバージェンスは s と t として分布 p と q を代入することにより得られる.
- u(•(p)) = p に注意. これは U-ダイバージェンスの持っているきれいな性質.

Example of *U* divergence

$$D_U(p,q)=\int U(\xi(q))-U(\xi(p))-p\{\xi(q)-\xi(p)\}$$

Kullback-Leibler (KL) divergence

$$U(t)=\exp(t)$$

$$D_{KL}(p,q)=\int\{q-p-p(\log q-\log p)\}=\int p\log\frac{p}{q}$$

Tsallis power divergence

$$U_\beta(t)=\frac{(1+\beta t)^{(1+\beta)/\beta}}{1+\beta}$$

$$D_\beta(p,q)=\int\{\frac{q^{\beta+1}-p^{\beta+1}}{\beta+1}-p\frac{(q^\beta-p^\beta)}{\beta}\}$$

Note $\lim_{\beta\to 0}U_\beta(t)=\exp(t)$

$\lim_{\beta\to 0}D_\beta(p,q)=D_{KL}(p,q)$

20

典型的な U-ダイバージェンス

- U を指数関数とすると KL ダイバージェンス.
- U-ダイバージェンスの中でも KL が一番いい. すなわちデータの分布が仮定された分布に exact に従うとすると U が指数であるというのが一番いい.
- Tsallis power ダイバージェンスもべき指数を使って構成できるので U-ダイバージェンスの一種. これはアウトライヤーの重みを 0 にするのでデータが汚れているときに有用.
- 現実においては仮定した統計モデルと現実とでは乖離があるので, それにあわせて Tsallis power ダイバージェンスの • を決める, という事が行われている.
- とはいえ, 実際には U の選び方はそんなに効かない. U の種類もそんなに知られていない.

U-エントロピー

U-クロスエントロピー　$C_U(p,q) = -E_p\{\xi(q)\} + \langle U(\xi(q))\rangle$

U-エントロピー　$H_U(p) = -E_p\{\xi(p)\} + \langle U(\xi(p))\rangle$

U-ダイバージェンス　$D_U(p,q) = C_U(p,q) - H_U(p)$

例題 1. Let $U(t) = \exp(t)$

Boltzmann-Shannonエントロピー　$H_U(p) = E_p(-\log p)$

例題 2. Let $U(t) = \frac{(t+\beta\cdot 1)^{\frac{1}{1+\beta}}}{1+\beta}$.

Tsallisエントロピー　$H_U(p) = E_p\{-\frac{p^\beta(X)-1}{\beta}\}$

ダイバージェンスの分解

- クロスエントロピーとエントロピーに分解される．
- $\bullet\cdot\bullet$ は積分を表す．
- E_p は p に関する期待値．

Geometric formula with D_U

$(g^{(U)}, \nabla^{(U)}, \nabla^{*(U)})$ s.t.　$g^{(U)}(X,Y) = -D_U(X|Y)$

$g^{(U)}(\nabla_X Y, Z) = -D_U(XY|Z) \quad (\forall Z \in \mathcal{X}(M))$

$g^{(D)}(\nabla^*_X Y, Z) = -D(Z|XY) \quad (\forall Z \in \mathcal{X}(M))$

$g^{(U)}_{ij}(\theta) = \int \frac{\partial}{\partial\theta^i} q(x,\theta) \frac{\partial}{\partial\theta^j} \xi(q(x,\theta))dx$

$\Gamma^{(U)}_{ij,k}(\theta) = \int \frac{\partial^2}{\partial\theta^i\partial\theta^j} q(x,\theta) \frac{\partial}{\partial\theta^k} \xi(q(x,\theta))dx$

$\Gamma^{*(U)}_{ij,k}(\theta) = \int \frac{\partial}{\partial\theta^k} q(x,\theta) \frac{\partial^2}{\partial\theta^i\partial\theta^j} \xi(q(x,\theta))dx$

$g^{(U)} = g \iff U = \exp$　(KL divergence)

$\nabla^{(U)} = \overset{m}{\nabla}$　$(\forall U)$　(Eguchi, 2005)

22

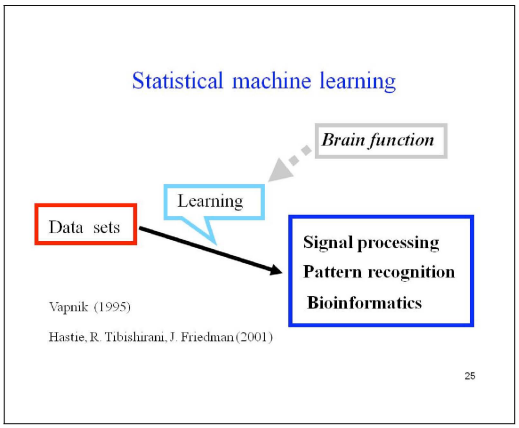
ダイバージェンスが作る幾何量である計量と組の接続を公式に基づいて計算

- パラメータを作って成分で計算．
- (任意の U に対し)• $\Gamma^{(U)}_{ij,k}$ は m-接続になっている．これは U-ダイバージェンスの特徴であり，統計的な方法に使うときに便利な性質である．

Application of
minimum divergence method
to Machine learning

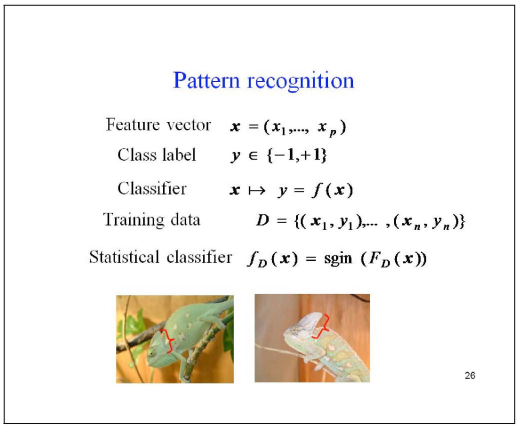
24

統計的な応用について



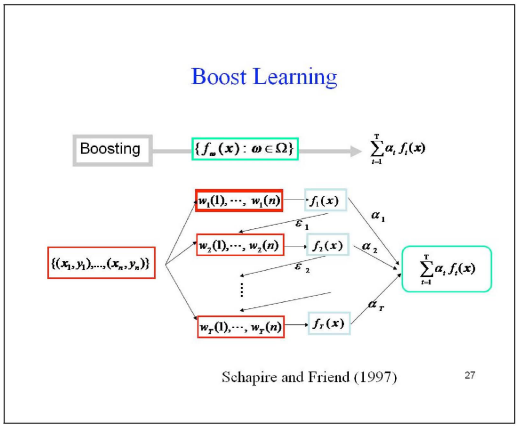
統計機械学習について

- ここ 4・5 年の統計の応用のはやり
- 「データをもとにして統計的な方法を行う」事を 学習する という (それ以前は「推定する」「検定する」).
- 統計機械学習はある意味，脳の機能を imitate して学習する.



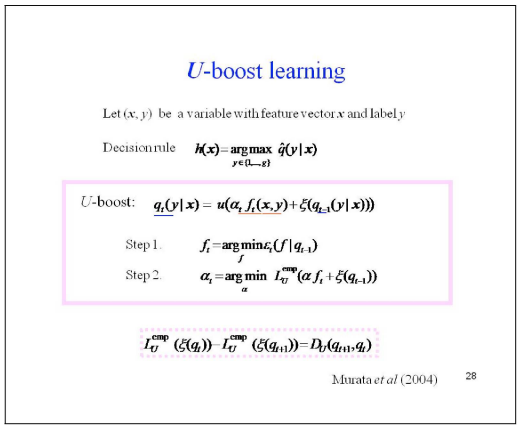
パターン認識について

- 識別子 f は $\mathbf{x}$ を見てラベルを予測する.
- 最終的に判別関数 $F_D(\mathbf{x})$ を作って，その符号によりラベルをわける.
- たとえばオスのエボシカメレオンがメスを獲得できるか，をラベルとすると識別子は頭頂のクレスト.



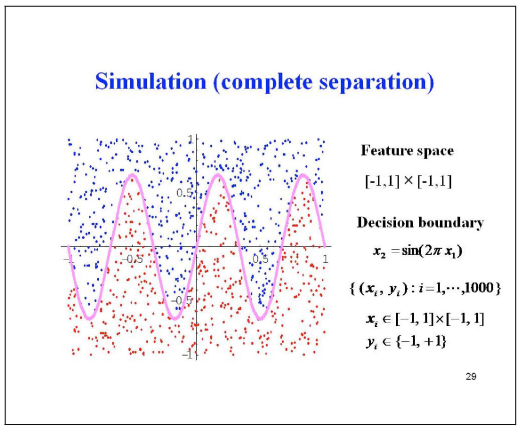
ブースト学習

- SVM (Support Vector Machine) に対抗.
- 識別子の辞書 $\{f_\cdot(x) : \cdot = 1, 2, \ldots\}$ を用意し，うまく多数決，すなわち重み $\cdot_t$ を付けて線形和をとることにより，よい統計分類を作る. このとき，辞書をたくさん用意すると，ロスを汎関数的に最小化できる.
- n 個の例題に対し，毎回重みを変えると，一番よかった識別子が毎回異なるので，それらの識別子を統合して多数決をとる.
- ポイントは重み $w_i(1), \ldots, w_i(n)$ の付け方.
  1. $w_1(1) = \ldots = e_1(n) = 1/n$.
  2. $f_1(\mathbf{x})$ が誤識別した例題のみ非常に大きくなるように重みを決定.
  3. 以降，Markov 的に重みを決定.

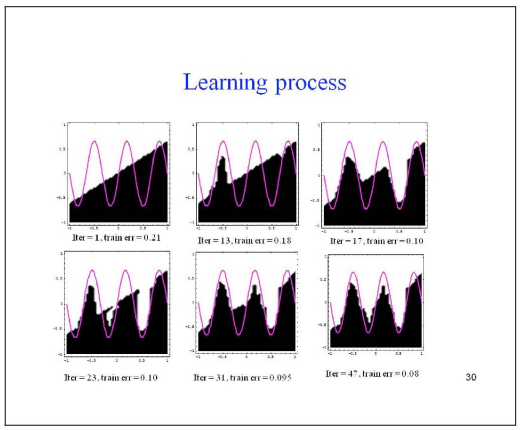


U-ダイバージェンスにブースト学習を応用

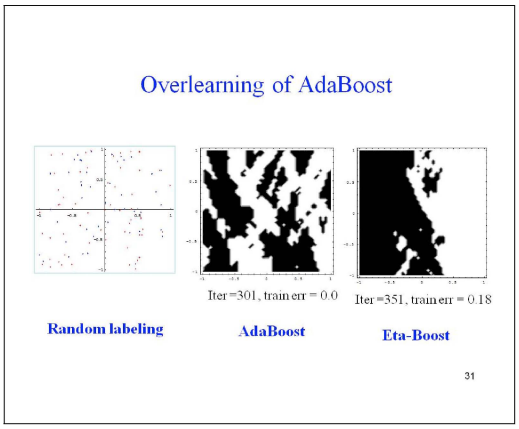
- $\bullet_t$ は U-ダイバージェンスが作るロスの gradient.
- $L_U^{emp}(\bullet(q_t))$ • $L_U^{emp}(\bullet(q_{t+1}))$ はステップ t とステップ t + 1 でのロスの減少幅．これが U-ダイバージェンスの 2 点間の距離と等しい，というピタゴラスの定理が成り立つ．



非線形な例：サイン波の上が +1, 下が • 1

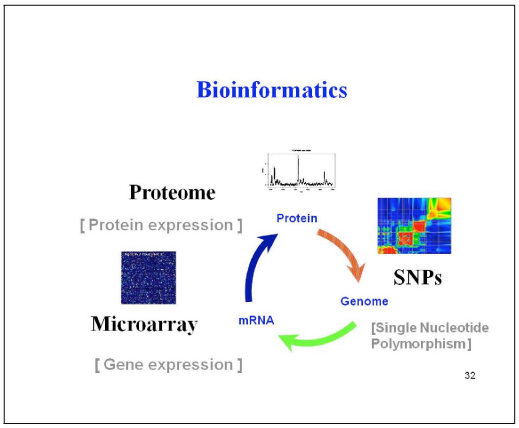


- 辞書は線形判別関数のみなので Step1 では線形.
- ブースト学習アルゴリズムが徐々に非線形性を学習する.



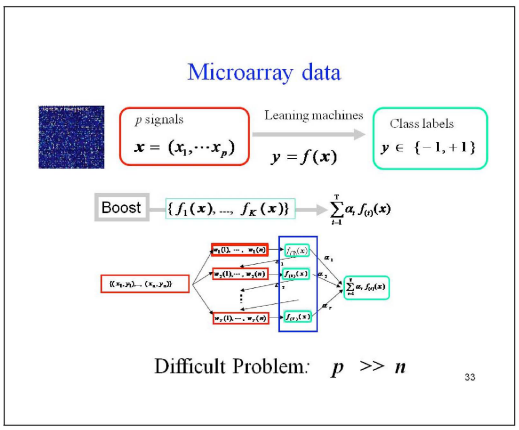
解がない場合，どういう学習をするか

- ランダムな点，ランダムなラベル．
- アダブーストでは中央図のようになんとかあわせようとする．過剰学習が起きる．
- U を少し変形すると右図のようになる．ひとかたまりになるので過剰学習に強い．
- ブーストの収束の判断は難しい．
  - 過剰学習が起きるので，回数を多くすればいいというものではない．
  - よくやられているのが情報量規準だが，ブースティングでまだよい情報量規準が出ていない．
  - 現在用いている方法:
    トレーニングデータを仮想的にトレーニングに使う分と残りにわけてテストする．これをすべての分け方で行い，テストエラーを小さくする回数を見つける．

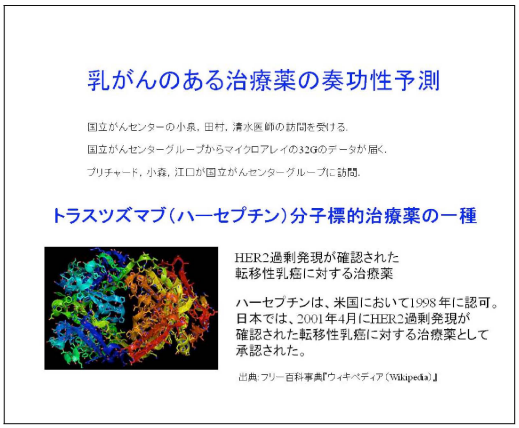


関連解析

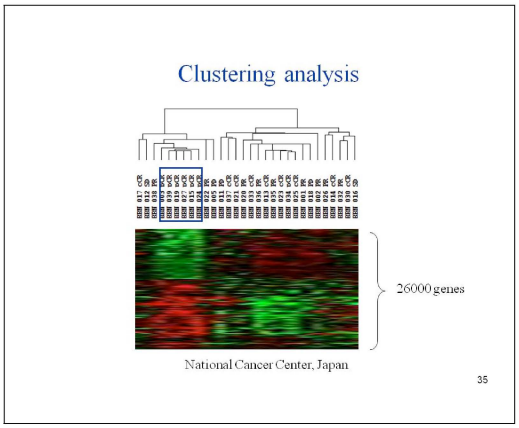
- バイオインフォマティクスの特別なタスク



- p 個の遺伝子の発現量．
- y は治療効果があるかどうか (テーラーメイド医療に関連)．
- n はサンプルサイズ．
- 医学的に解釈がつくような K を取り，ブーストアルゴリズムにかける．
- p は 2 万〜3 万，n は大体 100 以下．一方，p が 1 万だったら n が 100 万位だと統計的に有意な事ができる．したがってかなり無理なことをしている (p • n 問題)．エビデンスを他の外的なインフォメーション (パスウェイの知識，ネットワークの知識など) から持ってこないといけない．
- ブーストの辞書の選び方は発現の上位のものから．ただし，医学的に意味にないのも多いのでジーンリストに限って解析することもある．
- 各 f_i は一つの遺伝子・プローブしか見ていないので，線形な判別というよりステップワイズな判別．

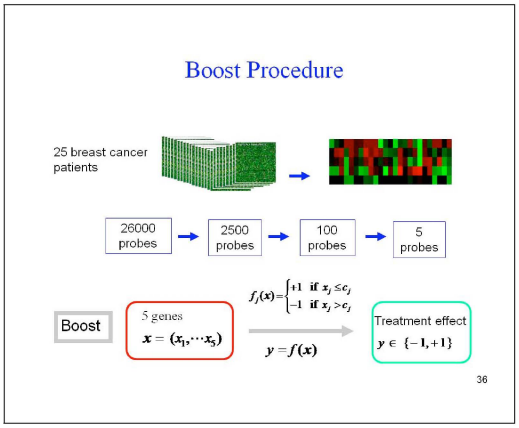


- 通常 2〜3 年目でデータの半分，5 年目ですべてのデータが届く．
- ハーセプチンはある乳癌の患者には全く効かない．あらかじめ遺伝子発現量を見て，効くかどうか予測できればテーラーメイド医療につながる．



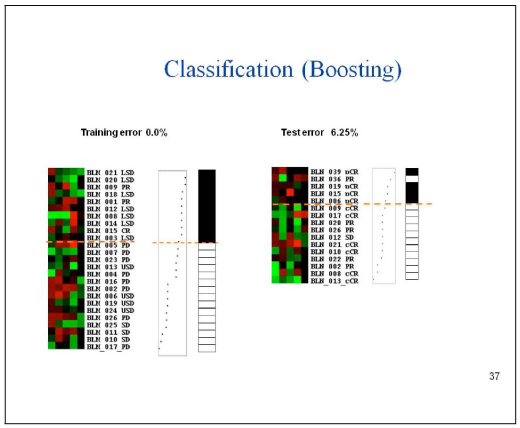
階層的クラスタリング

- 青く囲った部分は pCR，すなわち病理学的に完全に治った患者なので，非常にきれいなデータ (これが vague に出ているとどうしようもない)．
- 考慮しなかったが，本来ノイズにかなり影響されてしまうので実際に医療に使うには，たとえばクラスターをグループとしてもう少し安定な指標にする必要がある．
- アレイのサイエンスが進めば，ラーニングがスーパーバイズド・アンスーパーバイズドを行ったり来たり，ということで悩まなくてもよくなるのでは．



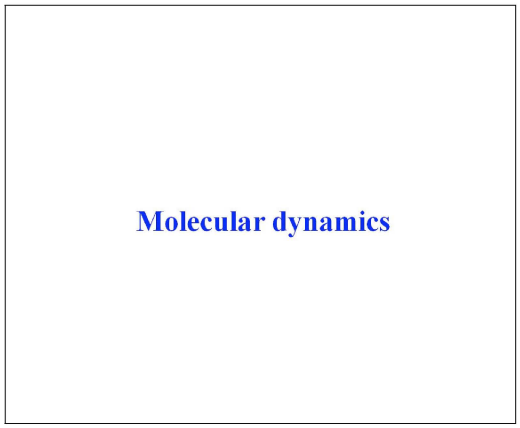
フィルタリング

- 発現が多く，データマイニングして遺伝子の機能がわかっているものを優先的に選択する．
- 最終的に 5 個の遺伝子での予測の成績が一番よかった (6 だと急にパフォーマンスが落ちる)．しかし，必ずしも機能がわかっている本命が出てきていない．統計的にフィットする予測がいいのか，それともアソシエーションがちゃんとつかまえられるのがいいのか，というのは別の問題．
- 予測に対する多重性の理論はできていないが，経験的に言ってこれだけうまくいく例はあまりない．
- 26000 個の遺伝子の中で選択した 5 個と似たような遺伝子もあわせて見る，という議論も現在進めている．

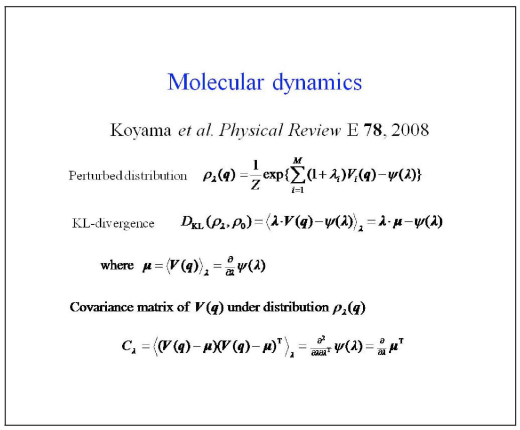


トレーニングデータとテストデータ

- アダブーストではいつでもトレーニングエラーを 0 にできる (ただし，その前にやめた方がいい場合もある).
- さらに追加のデータに対しては時間差を考えるなど改良する必要がある.



小山氏のメール・論文に関するコメント

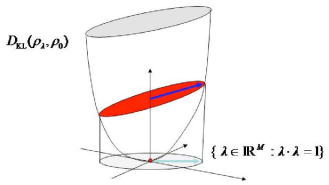


小山氏の論文

Max KL-divergence

Max KL-divergence on e-geodesic sphere

$$\max \{ D_{\mathrm{KL}}(\rho_\lambda, \rho_0) : \lambda \cdot \lambda = 1 \}$$



- 小山氏のしているのは制約付きの KL ダイバージェンスの最大化.
- 一方，統計で行うのは最小化．それが最尤法と等価，というのがもともとの KL ダイバージェンスの使い方.

PCA proposed by Koyama *et al.*

Max KL-divergence on e-geodesic sphere

$$\max \{ D_{\mathrm{KL}}(\rho_\lambda, \rho_0) : \lambda \cdot \lambda = 1 \}$$

Approximate ellipsoid to KL

$$D_{\mathrm{KL}}(\rho_\lambda, \rho_0) = \frac{1}{2} \lambda^{\mathrm{T}} C_0 \lambda + O(\| \lambda \|^3)$$

$$\arg\max \{ D_{\mathrm{KL}}(\rho_\lambda, \rho_0) : \lambda \cdot \lambda = 1 \} \approx \arg\max \{ \lambda^{\mathrm{T}} C_0 \lambda : \lambda \cdot \lambda = 1 \}$$

Max KL-divergence on e-geodesic sphere $\simeq$ PCA with C_0

- 二次近似して，C_0 を推定して PCA をすることにより KL ダイバージェンスの最大化と近似的に等価.

Iterative PCA?

$$\max \{ D_{\mathrm{KL}}(\rho_\lambda, \rho_0) : \lambda \cdot \lambda = 1 \}$$

Lagrange function $L(\lambda) = D_{\mathrm{KL}}(\rho_\lambda, \rho_0) - \frac{1}{2}\kappa(\lambda \cdot \lambda - 1)$

Gradient $\frac{\partial}{\partial \lambda} L(\lambda) = \frac{\partial}{\partial \lambda} \{ \lambda \cdot \mu - \psi(\lambda) - \frac{1}{2}\kappa(\lambda \cdot \lambda - 1) \} = C_\lambda \lambda - \kappa \lambda$

$$\lambda^* = \arg\max_{\lambda \cdot \lambda = 1} D_{\mathrm{KL}}(\rho_\lambda, \rho_0) = \arg\max_{\lambda \cdot \lambda = 1} \lambda^{\mathrm{T}} C_\lambda \lambda$$

Exact equivalence with PCA

Iteration algorithm for finding λ^*

$$\lambda_{t+1} \leftarrow \text{Principal-eigenvector}(\hat{C}_{\lambda_t})$$

$$\hat{C}_{\lambda_t} \lambda_{t+1} = \kappa \lambda_{t+1}$$

Exact に計算

- $\bullet^*$ を求めるために C$\bullet$ の PCA
- C と $\bullet$ は入れ子になってしまっている.
  - C を与えると第 1 固有値 $\bullet$ が定義される.
  - C は $\bullet$ の関数.
- 小山氏は摂動が小さいときには摂動前の情報から予測ができるので C_0 としていた.
- 別のアイデアとして，反復アルゴリズムを用いることにより exact な KL ダイバージェンスの最大化が得られるのではないか.

Boost Learning Algorithm?

Objective functional $L(V) = \langle V(q) \rangle_{\rho^{(V)}} - \psi(V) - \frac{1}{2}\kappa(\|V\|^2 - 1)$

where $\rho^{(V)}(q) = \rho_0(q)\exp\{V(q) - \psi(V)\}$

Dictionary of potential energy functions $F = \{f_\alpha(q) : \alpha \in A\}$

Boost learning algorithm

$V_t(q) \leftarrow V_{t-1}(q) + \lambda_t f_t(q)$

(i) $f_t = \arg\max\{\frac{\partial}{\partial\lambda} L(V_{t-1} + \lambda f)|_{\lambda=0} : f \in F\}$

(ii) $\lambda_t = \arg\max\{L(V_{t-1} + \lambda f_t) : \lambda \in \mathbb{R}\}$

$V(q) = \sum_{t=1}^{T} \lambda_t V_t(q)$

43

ブースト学習が可能

- あまり時間はかかからないのでは.

Max dual KL-divergence?

Max dual KL-divergence on e-geodesic sphere

$\max\{D_{KL}(\rho_0, \rho_\lambda) : \lambda \cdot \lambda = 1\}$

Lagrange function $L_\kappa(\lambda) = D_{KL}(\rho_0, \rho_\lambda) - \frac{1}{2}\kappa(\lambda \cdot \lambda - 1)$

Gradient $\frac{\partial}{\partial\lambda} L(\lambda) = \frac{\partial}{\partial\lambda}\{-\lambda \cdot \mu_0 + \psi(\lambda) - \frac{1}{2}\kappa(\lambda \cdot \lambda - 1)\} = -\mu_0 + \mu - \kappa\lambda$

$\mu_{\kappa^*} - \kappa^* \lambda_{\kappa^*} = \mu_0,\ \lambda_{\kappa^*} \cdot \lambda_{\kappa^*} = 1$

44

KL の非対称性より $\bullet_0$ と $\bullet$ を逆にする

- PCA の形にならない.
- 少し難しいかもしれないが停留点を見つければ計算できるのでは.
- どれくらい意味があるかはよくわからない.

Max Tsallis power divergence?

Tsallis power divergence

$D_\beta(\rho_\lambda, \rho_0) = \int \{\frac{\rho_0^{\beta+1} - \rho_\lambda^{\beta+1}}{\beta+1} - \rho_\lambda \frac{\rho_0^\beta - \rho_\lambda^\beta}{\beta}\}$

Tsallis entropy distribution

$\rho_\lambda(q) = \frac{1}{Z}\exp_\beta\{\sum_{i=1}^{M}(1 + \lambda_i)V_i(q) - \psi_\beta(\lambda)\}$

where $\exp_\beta(V) = (1 + \beta V)^{\frac{1}{\beta}}$

Max Tsallis power divergence on e-geodesic sphere ?

$\max\{D_\beta(\rho_\lambda, \rho_0) : \lambda \cdot \lambda = 1\}$

45

Tsallis power ダイバージェンスの最大化も考えられる