By Li xiaoran
www.sauron.online
xiaoranli@daum.net

Expected value-P1
Variance-P1
Covariance-p2
Covariance matrix-p2
High-dimensional Gaussian distribution-p2
Moment estimation-p2
統計的学習方法の紹介-p2
Generalization error bound-P5
Generative モデルと Discriminative モデル-P6
Classification problem-P6
Perceptron-P7
SGD（stochastic gradient descent）-P7
The dual form of the perceptron -P8
k-NN（k-Nearest Neighbor）-P9
Naive Bayes-P10
MLE(maximum likelihood estimation)-P11
Bayesian estimation-P11
Maximize posterior probability Proof-P12
Maximum likelihood estimation VS Bayesian estimation -P12
Decision tree（if-then 規則）-P13
Entropy-P14
Conditional entropy-P14
Information gain-P14
Logistic regression & Maximum entropy-P16
Binomial logistic regression model-P16
Maximum entropy model-P17
Intuition of Equality Constraint-P18
Lagrangian Dualization-P19
KKT-P19
MLE of Maximum entropy mode-P19
Proof of improved iterative scaling (IIS)-P20
SVM-P21
Proof : SVM Dual problem-P23
The uniqueness of hyperplane existence-P24
Boosting Adaboost-P25
Expectation Maximization (EM)-P26
Gaussian mixture model (GMM)-P27
Hidden Markov model (HMM)-P28
Conditional random field （CRF）-P30
Newton Method-P32

Expected value

1） $E(C) = C$

2） $E(X+Y) = E(X) + E(Y)S$

3） Independence：$E(XY) = E(X)E(Y)$

Proof： $f_{XY}(x,y) = \int p(x,y)dx \int p(x,y)dy$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_{XY}(x,y)dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_X(x) f_Y(y)dxdy$$
$$= \int_{-\infty}^{\infty} x f_X(x)dx \int_{-\infty}^{\infty} y f_Y(y)dy$$

Variance

1） $D(X) = Var(X) = E\{[X - E(X)]^2\} = E(X^2) - E(X)^2$

$$= E\{X^2 - 2XE(X) + [E(X)]^2\}$$
$$= E(X^2) - 2E[XE(X) + E[E(X)^2]$$
$$= E(X)^2 - 2E(X)E(X) + E(X)^2$$
$$= E(X^2) - E(X)^2$$

2） $D(C) = 0$

3） $D(CX) = C^2 D(X)$

$$= E[(CX)^2] - E^2(CX)$$
$$= E[C^2X^2] - [CE(X)]^2$$
$$= C^2E(X)^2 - C^2E^2(X)$$

4） $D(X+C) = D(X)$

$$= E\{[(X+C) - E(X+C)]^2\}$$
$$= E\{[X + C - (E(X) - C)]^2\}$$
$$= E\{X - E(X)^2\}$$

5） $D(X+Y) = D(X) + D(Y) + 2E\{[X - E(X)][Y - E(Y)]\}$

$$= E\{[(X+Y) - E(X+Y)]^2\}$$
$$= E\{[X - E(X)]^2\} + E\{[Y - E(Y)]^2\} + 2E\{[X - E(X)][Y - E(Y)]\}$$

6） Independence：$D(X+Y) = D(X) + D(Y)$

$$= \cdots + 2E\{[X - E(X)][Y - E(Y)]\}$$
$$= \cdots + 2E\{XY - XE(Y) - YE(X) + E(X)E(Y)\}$$
$$= \cdots + 2\{E(XY) - E(X)E(Y) - E(Y)E(X) + E(X)E(Y)\}$$
$$= \cdots + 2\{E(XY) - E(X)E(Y)\}$$

Covariance

$Cov(X,Y) = E\{[X - E(X)][Y - E(Y)]\}$

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$$

1） $Cov(aX, bY) = abCov(X,Y)$

2） $Cov(X_1 + X_2, Y) = Cov(X_1, Y) + Cov(X_2, Y)$

Covariance matrix

$$c_{11} = E\{[X_1 - E(X_1)]^2\}$$

$$c_{12} = E\{[X_1 - E(X_1)][X_2 - E(X_2)]\}$$
$$c_{21} = E\{[X_1 - E(X_1)][X_2 - E(X_2)]\}$$
$$c_{22} = E\{[X_2 - E(X_2)]^2\}$$

High-dimensional Gaussian distribution

$$r.v\, X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}; \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{pmatrix}$$

$$\frac{1}{(2\pi)^{\frac{n}{2}}(\det C)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(X-\mu)^T C^{-1}(X-\mu)\right\}$$

Moment estimation

$$\mu_l = E(X^l) = \int_{-\infty}^{\infty} x^l f(x;\theta_1,\theta_2,\dots,\theta_n)$$

$$A_l = \frac{1}{n}\sum_{i=1}^{n} X_i^l$$

$$\mu_l = A_l$$

例：

$$\begin{cases} \mu_1 = E(X) = \mu \\ \mu_2 = E(X^2) = D(X) + [E(X)]^2 = \sigma^2 + \mu^2 \end{cases} = \begin{cases} \mu = \mu_1 \\ \sigma^2 = \mu_2 - \mu_1^2 \end{cases}$$

$$\text{Moment estimation:} \begin{cases} \hat{\mu} = A_1 = \bar{X} \\ \hat{\sigma}^2 = A_2 - A_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \end{cases}$$

1. 統計的学習方法の紹介

1, 1 方法=モデル+戦略+アルゴリズム

1. 1. 1 モデル

Decision function

$$\mathcal{F} = \{P|Y = f_\theta(X), \theta \in Parameter\ Space^n\}$$
$$Y = a_0 + a_1 X\,; \quad \theta = (a_0, a_1)^T$$

Conditional probability

$$\mathcal{F} = \{P|Y = P_\theta(Y|X), \theta \in Parameter\ Space^n\}$$
$$Y \sim N(a_0 + a_1 X, \sigma^2)$$

1. 1. 2 戦略＝Decision Loss 或は Conditional probability Loss の選択

Decision Loss ：

0-1 loss function: $L\big(Y, f(X)\big) = \begin{cases} 1, & Y \neq f(X) \\ 0, & Y = f(X) \end{cases} \quad \rightarrow$

$Classification(discrete)$

quadratic loss function: $L(Y,f(X)) = (Y-f(X))^2 \quad \rightarrow$
$Regression(continuous)$ 強い

absolute loss function: $L(Y,f(X)) = |Y-f(X)| \quad \rightarrow$
$Regression(continuous)$

Conditional probability Loss : ≠0

logarithmic/log-likelihood loss function: $L(Y,P(Y|X)) = -\log P(Y|X)$

Expected Loss 或は Risk function：

$$R_{exp}(f) = E_p[L(Y,f(X))] = \int_{x\times y} L(y,f(x))P(x,y)dxdy$$

Empirical Risk/Loss：
S.t: $data = \{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$

$$R_{emp}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i,f(x_i))$$

ERM（empirical risk minimization）：
例えば：Loss function＝Log-likelihood loss function の時 ERM ⇔ MLE（maximum likelihood estimation）

$$\underset{f\in\mathcal{F}}{minimize}\frac{1}{N}\sum_{i=1}^{N} L(y_i,f(x_i))$$

Regularization ⇔ SRM (structural risk minimization)：over-fitting の出ないため

Empirical Risk/Loss＋Regularizer/Penalty term：$J(f)$がモデルの複雑を反映し、fが複雑になるほど$J(f)$は大きくなる。ラムダは 0 以上。

$$R_{srm}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i,f(x_i)) + \lambda J(f)$$

例えば：Bayesian estimation の中に MAP（maximum posterior probability estimation）が SRM の一つ。もし Loss function は Log-likelihood loss function だ、Regularizer/Penalty term は Priori probability だの時 SRM ⇔ MAP。

1.1.3 アルゴリズム= Optimization problem

1.2 モデルの評価と選択

1.2.1 Training error と **Test error** 両者の Loss function 必ず同じではない、でも同じほうがいいと思う。

Error rate と Accuracy：I が Indicator function だ（$y \neq f(x)$時 1 になる、他の時 0 になる）$e_{test} + r_{test} = 1$

$$e_{test} = \frac{1}{N'}\sum_{i=1}^{N'} I\left(y_i \neq \hat{f}(x_i)\right)$$

$$r_{test} = \frac{1}{N'}\sum_{i=1}^{N'} I\left(y_i \neq \hat{f}(x_i)\right)$$

1.2.2 Over-fitting

S.t: $data = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$

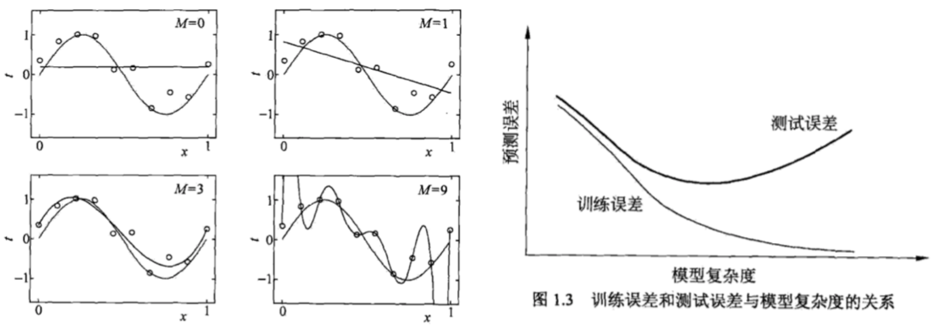
$f_M(x, w) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j$

$Loss(w) = \frac{1}{2}\sum_{i=1}^{N}(f(x_i, w) - y)^2$

$$L(w) = \frac{1}{2}\sum_{i=1}^{N}\left(\sum_{j=0}^{M} w_j x_i^j - y_i\right)^2$$

“w”の部分導関数を求めさせる：

$$w_j = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^{j+1}}, j = 0,1,2,\cdots,M$$



図： $y = \sin\frac{x}{2\pi} + \varepsilon$

1.2.3 Regularization/Penalty term

Loss functionに従ってRegularizerを使う。例えば：Regresstic problemに対してLoss functionがquadratic loss functionになり、L2のRegularizerがよく使えられる：Occam’s razor：RegularizerはPrioriを見る時、複雑のモデルほうが複雑のPrioriである

$$L(w) = \frac{1}{N}\sum_{i=1}^{N}(f(x_i; w) - y_i)^2 + \frac{\lambda}{2}\|w\|^2$$

L1のRegularizerになれば：

$$L(w) = \frac{1}{N}\sum_{i=1}^{N}(f(x_i; w) - y_i)^2 + \lambda\|w\|_1$$

Generalization ability ⇔ Expected Loss或はRisk function

1.3 Generalization error bound:

$$R(f) = E\big[L\big(Y, f(X)\big)\big]$$

$$\hat{R}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i))$$

$F = \{f_1, f_2, \cdots, f_d\}$; $N = \#(samples)$

$$R_{\text{Expected Risk}}(f) \le \hat{R}_{\text{empirical risk}}(f) + \varepsilon(d, N, \delta)$$

$$\varepsilon(d, N, \delta) = \sqrt{\frac{1}{2N}\left(\log d + \log\frac{1}{\delta}\right)}$$

Proof: $\boldsymbol{R(f) \le \hat{R}(f) + \sqrt{\frac{1}{2N}\left(\log d + \log\frac{1}{\delta}\right)}}$

Hoeffding inequality :

s.t $ES_n = E(S_n)$, $S_n = \sum_{i=1}^{n} X_i$, $X_i \in [a, b], t > 0$

$$P(S_n - ES_n \ge t) \le \exp\left(\frac{-2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

Intuition of Hoeffding inequality

$$\bar{X}_n = \frac{S_n}{n}\ ; \quad E(\bar{X}_n) = \frac{ES_n}{n}$$

$$P(\bar{X}_n - E(\bar{X}_n) \ge t) = P(S_n - ES_n \ge nt) \le \exp\left(\frac{-2nt^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) = e^{-n}$$

$$n = \infty\ ;\ P(\bar{X}_n - E(\bar{X}_n) \ge t) \to 0$$

Substituting $R(f) = E\big[L\big(Y, f(X)\big)\big]$; $\hat{R}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i))$

into $P(E\bar{X}_n - \bar{X}_n \ge t) \le \exp\left(\frac{-2nt^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$

$$P\big(R(f) - \hat{R}(f) \ge t\big) \le \exp\left(\frac{-2Nt^2}{N}\right) = \exp(-2Nt^2)$$

$$P\big(\exists f \in F, R(f) - \hat{R}(f) \ge t\big) = P\big(\ R(f_1) - \hat{R}(f_1) \ge t, \cup, \cdots, \cup, R(f_d) - \hat{R}(f_d) \ge t\big)$$

$$\le \sum_{i} P\big(\ R(f_i) - \hat{R}(f_i) \ge t\big)$$

$$\le d\exp(2Nt^2)$$

$$P\big(\forall f \in F, R(f) - \hat{R}(f) \le t\big) \ge 1 - d\exp(2Nt^2) = \delta$$

$$t = \sqrt{\frac{1}{2N}(\log d + \log \delta)}$$

Proof completed

1.4 Generative モデルと Discriminative モデル

教師あり学習の中に Generative/ Discriminative approach がある

Generative approach：P(X,Y)の分布を学習し、そしてこの分布に基づいて X
を予測する。例えば：Naive Bayes；HMM

$$P(Y|X) = \frac{P(X,Y)}{P(X)}$$

Discriminative approach：直接に Decision function を学習する、例えば：K-NN，Perceptron，Decision tree，Logistic regression，Maximum entropy model，Support Vector Machines（SVM），Conditional random field（CRF）

$$f(X) \text{ 或 } \mathrm{P(Y|X)}$$

1.5 Classification problem

		真の結果	
		正	負
予測結果	正	TP	FP
	負	FN	TN

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{FP + TN}$$

$$\frac{2}{F - measure} = \frac{1}{Rcall} + \frac{1}{Precision}$$

$$F - measure = \frac{2Recall \cdot Precision}{Recall + Precision}$$

2. Perceptron

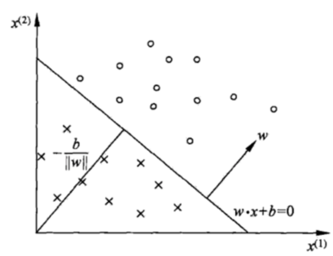


图 2.1 感知机模型

Linear classification model/Linear classifier を属する。$\{f|f(x) = w \cdot x + b\}$

S. t:$x \in \mathcal{X}_{(Input\ space(Feature\ space))} \subseteq \mathbb{R}^n; y \in \mathcal{Y}_{(Output\ space)} = \{+1, -1\}$

$$sign(x) = \begin{cases} +1, & x \geq 0 \\ -1, & x < 0 \end{cases}$$
$$f(x) = sign(w \cdot x + b)$$

Separating Hyperplane：$\vec{w} \cdot \vec{x} + \mathrm{b} = 0$；Normal vector：$\vec{w}$；Intercept：b

Proof of Normal vector：$\vec{w}$

$$\begin{cases} \vec{w} \cdot \vec{x_1} + \mathrm{b} = 0 & ① \\ \vec{w} \cdot \vec{x_2} + \mathrm{b} = 0 & ② \end{cases} \overset{①-②}{\Longrightarrow} \begin{cases} \vec{w}(\vec{x_1} - \vec{x_2}) = 0 \\ \vec{x_1} - \vec{x_2} = \overrightarrow{x_1 x_2} \end{cases}$$

x_i から Separating Hyperplane：S への距離：

$$\frac{1}{\|w\|}|w \cdot x_i + b|$$

誤分類の x_i から Separating Hyperplane：S への総距離：

$$-\frac{y_i(w \cdot x_i + b)}{\|w\|} = \frac{|w \cdot x_i + b|}{\|w\|}$$

S. t: $data = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$；誤分点の $x_i\ in\ M$

$$\min_{w,b} L(w, b) = -\sum_{x_i \in M} y_i(w \cdot x_i + b)$$

2.1 SGD（stochastic gradient descent）

$$\nabla_w L(w, b) = -\sum_{x_i \in M} y_i x_i$$

$$\nabla_b L(w, b) = -\sum_{x_i \in M} y_i$$

①誤分点(x_i, y_i)ランダームに選択して

$while$ (Training data)do

②Training data からw_0, b_0を選択して

③ if： $y_i(w \cdot x_i + b) \leq 0$ ， wとbを更新する：Learning rate η: $(0 < \eta \leq 1)$

$$w \leftarrow w + \eta y_i x_i$$
$$b \leftarrow w + \eta y_i$$

④ $else$：

②

$return\ w, b$

The dual form of the perceptron ：

The increment of w, b with respect to (x_i, y_i) is $\alpha_i y_i x_i$, $\alpha_i y_i$

$\alpha_i = n_i \eta$; $n = \#(iterations)$

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i$$

$$b = \sum_{i=1}^{N} \alpha_i y_i$$

$$f(x) = sign\left(\sum_{j=1}^{N} \alpha_j y_j x_j \cdot x + b\right); \ \alpha = (\alpha_1 \alpha_2, \cdots, \alpha_n)^T$$

$while$ (Training data)do

①$\alpha \leftarrow 0, b \leftarrow 0$

②Training data からw_i, b_iを選択して

③if： $y_i\left(\sum_{j=1}^{N} \alpha_j y_j x_j \cdot x + b\right) \leq 0$ ， wとbを更新する

$$\alpha_i \leftarrow \alpha_i + \eta$$
$$b \leftarrow b + \eta y_i$$

④ $else$：

②

$return\ \alpha, b$

2.2 Proof of convergence（Novikoff）

S.t Data: Linearly separable; $x \in \mathcal{X} = \mathbb{R}^n; y \in \mathcal{Y} = \{+1, -1\}; i = 1,2,\cdots,N;$

$\widehat{w} = (w^T, b)^T; \hat{x} = (x^T, 1)^T \rightarrow \hat{x} \in \mathbb{R}^{n+1}, \widehat{w} \in \mathbb{R}^{n+1}, \widehat{w} \cdot \hat{x} = w \cdot x + b$

Exist: $\|\widehat{w}_{opt}\| = 1\ to\ \widehat{w}_{opt} \cdot \hat{x} = w_{opt} \cdot x_i + b_{opt} = 0;\ \gamma > 0$

$$y_i(\widehat{w}_{opt} \cdot \hat{x}_i) = y_i(w_{opt} \cdot x_i + b_{opt}) \geq \gamma$$

S.t $R = \max_{1 \leq i \leq N} \|\hat{x}_i\|$; Training data 誤分類数 k

$$k \leq \left(\frac{R}{\gamma}\right)^2$$

サンプルを完全に分離できるハイパープレーンがある場合は

$$w \cdot x + b = 0$$
$$w_{opt} \cdot x + b_{opt} = 0$$

再定義w_{opt} , b_{opt}

$$\widehat{w}_{opt} = \left(w_{opt}^T, b_{opt}\right)^T$$

$$\hat{x} = (x^T, 1)^T$$

$$\widehat{w}_{opt} \cdot \hat{x} = w_{opt}^T \cdot x + b_{opt}$$

$$\therefore \quad w_{opt} \cdot x + b_{opt} = 0 \leftrightarrow \hat{w}_{opt} \cdot \hat{x} = 0$$

(1) $\exists \|\hat{w}_{opt}\| = 1, \hat{w}_{opt} \cdot \hat{x} = 0$

$Proof$: $\exists \gamma > 0, y_i(\hat{w}_{opt} \cdot \hat{x}) \geq \gamma$

$$y_i(\hat{w}_{opt} \cdot \hat{x}) > 0$$

$$\gamma = \min_i y_i(\hat{w}_{opt} \cdot \hat{x})$$

(2) $s.t\ R = \max\|\hat{x}_i\|,\ k \leq \left(\frac{R}{\gamma}\right)^2$

誤分類点が見つかるたびに、wが修正される。$k = \#(修正)$

1. $proof \quad \hat{w}_k \cdot \hat{w}_{opt} \geq k\eta\gamma$ （角度が段々近くになる）

$$\begin{aligned}\hat{w}_k \cdot \hat{w}_{opt} &= (\hat{w}_{k-1} + \eta y_i \hat{x}_i) \cdot \hat{w}_{opt} \\ &= \hat{w}_{k-1} \cdot \hat{w}_{opt} + \eta y_i \hat{x}_i \cdot \hat{w}_{opt} \\ &\geq \hat{w}_{k-1} \cdot \hat{w}_{opt} + \eta\gamma \\ &\geq \hat{w}_{k-2} \cdot \hat{w}_{opt} + \eta\gamma + \eta\gamma \\ &\vdots \\ &\geq \hat{w}_0 \cdot \hat{w}_{opt} + k\eta\gamma\end{aligned}$$

$$\because \hat{w}_0 = (0, \cdots, 0)^T \quad \therefore \geq k\eta\gamma$$

2. $proof \quad \|\hat{w}_k\|^2 \leq k\eta^2 R^2$ （長さが限界をある）

$$\begin{aligned}&\|\hat{w}_k\|^2 = \|\hat{w}_{k-1} + \eta y_i \hat{x}_i\|^2 \\ &= \|\hat{w}_{k-1}\|^2 + 2\,\eta y_i \hat{w}_{k-1} \cdot \hat{x}_i + \eta^2\|\hat{x}_i\|^2 \\ &\leq \|\hat{w}_{k-1}\|^2 + \eta^2 R^2 \\ &\vdots \\ &\leq \|\hat{w}_0\|^2 + k\eta^2 R^2 \\ &\leq k\eta^2 R^2\end{aligned}$$

$(1) + (2)$: Cauchy Inequality

$$\hat{w}_k \cdot \hat{w}_{opt} \leq \|\hat{w}_k\| \cdot \|\hat{w}_{opt}\|$$

$$k\eta\gamma \leq \hat{w}_k \cdot \hat{w}_{opt} \leq \|\hat{w}_k\| \leq \sqrt{k\eta^2 R^2}$$

$$k\eta\gamma \leq \sqrt{k\eta^2 R^2}$$

$$k \leq \left(\frac{R}{\lambda}\right)^2$$

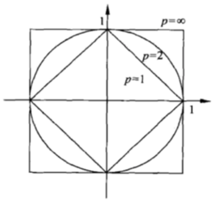
3. k-NN（k-Nearest Neighbor）

3.1 Distance measure: $L_p(x_i, x_j) = \left(\sum_{l=1}^{n} \left|x_i^{(l)} - x_j^{(l)}\right|^p\right)^{\frac{1}{p}}$

S.t $P=2$の時：Euclidean distance：$L_2(x_i,x_j)=\left(\sum_{l=1}^{n}\left|x_i^{(l)}-x_j^{(l)}\right|^2\right)^{\frac{1}{2}}$

S.t $P=1$の時：Manhattan distance：$L_1(x_i,x_j)=\sum_{l=1}^{n}\left|x_i^{(l)}-x_j^{(l)}\right|$

S.t $P=\infty$の時：$L_\infty(x_i,x_j)=$，$\max_l\left|x_i^{(l)}-x_j^{(l)}\right|$



4. Naive Bayes

Input space eigenvector： $x\in\mathcal{X}\subseteq\mathbb{R}^n$
Output space class label： $y\in\mathcal{Y}=\{c_1,c_2,\cdots,c_k\}$
Joint probability distribution：$P(X,Y)$
Training data：$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$
Priori distribution：$P(Y=c_k),k=1,2,\cdots,K$
Conditional probability distribution：

$$P(X=x|Y=c_k)=P\left(X^{(1)}=x^{(1)},\cdots,X^{(n)}=x^{(n)}\middle|Y=c_k\right),k=1,2,\cdots,K$$

Conditional independence hypothesis：$\prod_{j=1}^{n}P(X^{(1)}=x^{(1)}|Y=c_k)$

Maximize posterior probability：

$$P(Y=c_k|X=x)=\frac{P(X=x|Y=c_k)\cdot P(Y=c_k)}{\sum_k P(X=x|Y=c_k)\cdot P(Y=c_k)}$$
$$=\frac{P(Y=c_k)\prod_j P\left(X^{(j)}=x^{(j)}\middle|Y=c_k\right)}{\sum_k P(Y=c_k)\prod_j P\left(X^{(j)}=x^{(j)}\middle|Y=c_k\right)},k=1,2,\cdots,k$$

Then:

$$y=f(x)=\underset{c_k}{argmax}\frac{P(Y=c_k)\prod_j P\left(X^{(j)}=x^{(j)}\middle|Y=c_k\right)}{\sum_k P(Y=c_k)\prod_j P\left(X^{(j)}=x^{(j)}\middle|Y=c_k\right)}$$
$$=\underset{c_k}{argmax}\,P(Y=c_k)\prod_j P\left(X^{(j)}=x^{(j)}\middle|Y=c_k\right)$$

$$P(Y=C_1)=\frac{\#\{y_i=c_1\}}{N}$$
$$P(Y=C_2)=\frac{\#\{y_i=c_2\}}{N}$$

$$P\left(X^{(1)}=x_1^{(1)}\middle|Y=c_1\right)=\frac{\#\{y_i=c_1,X^{(1)}=x_1^{(1)}\}}{\#\{y_i=c_1\}}$$

4.1 MLE(maximum likelihood estimation)：MLE を使って likelihood と priori を予測する：

$$P(Y=c_k)=\frac{\sum_{i=1}^{N}I(y_i=c_k)}{N},k=1,2,\cdots,K$$

$$P\left(X^{(j)}=a_{jl}\middle|Y=c_k\right)=\frac{\sum_{i=1}^{N}I\left(x_i^{(j)}=a_{ij},y_i=c_k\right)}{\sum_{i=1}^{N}I(y_i=c_k)}$$

$Y\in\{c_1,c_2,\cdots,c_k\}\longrightarrow\{\theta_1,\theta_2,\cdots,\theta_k\}$; $\sum_{i=1}^{k}\theta_i=1$

$$P(Y=y|\theta)=\theta^{I(y=c_1)}\cdot\theta^{I(y=c_2)}\dots\theta^{I(y=c_k)}$$

$$P(y_1,y_2,\cdots,y_N|\theta)=P(y_1|\theta)P(y_2|\theta)\cdots P(y_N|\theta)$$

$$=\theta_1^{m_1}\cdot\theta_2^{m_2}\cdots\theta_k^{m_k}\ ;\ m_i=\#(c_i)\ ;\ \sum m_i=N$$

$$\Longrightarrow\max\ln P(y_1,y_2,\cdots,y_N|\theta)=m_1\ln\theta_1+m_2\ln\theta_2+\cdots+m_k\ln\theta_k$$

$s.t.\theta_1+\theta_2+\cdots+\theta_k=1$ lagrangian

$$L(\theta,\lambda)=\max_{\theta_1,\cdots,\theta_k}m_1\ln\theta_1+m_2\ln\theta_2+\cdots+m_k\ln\theta_k+\lambda(\theta_1+\theta_2\cdots+\theta_k-1)$$

$$\frac{\partial L(\theta,\lambda)}{\partial\lambda}:-\frac{m_1+m_2+\cdots+m_k}{\lambda}=1\Longrightarrow\lambda=-m_1+m_2+\cdots+m_k=-N$$

$$\frac{\partial L(\theta,\lambda)}{\partial\theta_i}:\frac{m_i}{\theta_i}+\lambda=0\Longrightarrow\theta_i=-\frac{m_i}{\lambda}=\frac{m_i}{N}\ ;\ i=1,2,\cdots,K$$

4.1.1 Bayesian estimation：MLE を使う時もし確率がゼロになれば Posterior probability に影響される。だから Bayesian estimation を使う。

Likelihood：$\lambda=0$ の時 Bayesian estimation ⇔ MLE；$\lambda=1$ になれば Laplace smoothing

$$P\left(X^{(j)}=a_{jl}\middle|Y=c_k\right)=\frac{\sum_{i=1}^{N}I\left(x_i^{(j)}=a_{ij},y_i=c_k\right)+\lambda}{\sum_{i=1}^{N}I(y_i=c_k)+S_j\lambda}$$

Priori：$Y(\theta_1,\theta_2,\cdots,\theta_k)$ $\sum=1$ $\sim\text{Dirichlet (Dirichlet} = \text{beta}^{\text{n}})$

$$\pi(\theta)\frac{\Gamma(\alpha_1+\alpha_2+\cdots\alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_k)}\theta_1^{\alpha-1}\theta_2^{\alpha-1}\cdots\theta_K^{\alpha-1}$$

$$P(Y=c_k)=\frac{\sum_{i=1}^{N}I(y_i=c_k)+\lambda}{N+K\lambda}$$

$$p(\theta|y_1,\cdots,y_n)=\frac{p(\theta,y_1,\cdots,y_n)}{p(y_1,\cdots,y_n)}$$

$$\propto p(\theta)p(y_1,\cdots,y_n|\theta)$$

$$\propto\theta_1^{\alpha-1}\theta_2^{\alpha-1}\cdots\theta_K^{\alpha-1}\cdot\theta_1^{m_1}\theta_2^{m_2}\cdots\theta_K^{m_3}$$

$$\propto\theta_1^{m_1+\alpha-1}\theta_2^{m_2+\alpha-1}\cdots\theta_K^{m_3+\alpha-1}\Longrightarrow(\text{Dirichlet})\,;e^{-\theta^2+a\theta+b}$$

$$\Longrightarrow(\text{Normal})$$

$$=\frac{\Gamma(m_1+\alpha-1+m_2+\alpha-1+\cdots m_3+\alpha-1)}{\Gamma(m_1+\alpha-1)\Gamma(m_2+\alpha-1)\cdots\Gamma(m_3+\alpha-1)}\theta_1^{m_1+\alpha-1}\theta_2^{m_2+\alpha-1}\cdots\theta_K^{m_3+\alpha-1}$$

$$\theta_i=\frac{m_i+\alpha-1}{\sum m_k+\alpha-1}$$

$$= \frac{m_i + \alpha - 1}{N + K(\alpha - 1)}$$

4.1.2 Maximize posterior probability Proof

$$L(Y, f(x)) = \begin{cases} 1, & Y = f(x) \\ 0, & Y \neq f(x) \end{cases}$$

Proof ： $\boldsymbol{f(x) = \underset{c_k}{argmax} P(Y = c_k|x)}$

$$\min E\, L(Y, f(x)) = \sum_Y \sum_x [L(Y, f(x)) P(x, Y)]$$

$$= \sum_x \left[\sum_Y L(Y, f(x)) P(Y|x) \right] P(x)$$

$$\propto \min \sum_Y L(Y, f(x)) P(Y|x)$$

$$\propto \min \sum_{c_k} L(Y = c_k, f(x)) P(Y = c_k|x)$$

$$\propto \min \sum_{c_k} I(f(x) \neq c_k) P(Y = c_k|x)$$

$$\propto \min \sum_{c_k} [1 - I(f(x) \neq c_k)]\, P(Y = c_k|x)$$

$$\propto \min \sum_{c_k} P(Y = c_k|x) - \sum_{c_k} I(f(x) \neq c_k) \cdot P(Y = c_k|x)$$

$$\propto \min 1 - \sum_{c_k} I(f(x) \neq c_k) \cdot P(Y = c_k|x)$$

$$\propto \max \sum_{c_k} I(f(x) \neq c_k) \cdot P(Y = c_k|x)$$

$$\therefore \quad f(x) = \underset{c_k}{argmax} P(Y = c_k|x)$$

4.2 Maximum likelihood estimation VS Bayesian estimation ：

s.t $X_i = \begin{cases} 1, & postive \\ 0, & negative \end{cases}$ ； $X_i \sim b(1, \theta)$ ； $P(X = x) = \theta^x (1 - \theta)^{1-x}$

$$L(\theta) = P(x_1 = x_1|\theta) \cdots P(x_n = x_n|\theta)$$

$$= \prod_{i=1}^{n} \theta^{x_i} (1 - \theta)^{1-x_i}$$

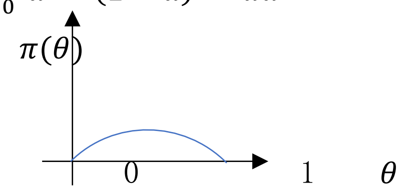
MLE ：

$$\max \ln L(\theta) = \sum [\ln \theta^{x_i} + \ln(1-\theta)^{1-x_i}]$$
$$= \sum \ln \theta + (n - \sum x_i) \ln(1-\theta)$$
$$\frac{\partial \ln L(\theta)}{\partial \theta} = \frac{\sum x_i}{\theta} - \frac{n \cdot \sum x_i}{1-\theta} = 0$$
$$\hat{\theta} = \frac{\sum x_i}{n}$$

BE:

S.t Priori : θ~Bata(α,β) : PDF

$$\beta(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

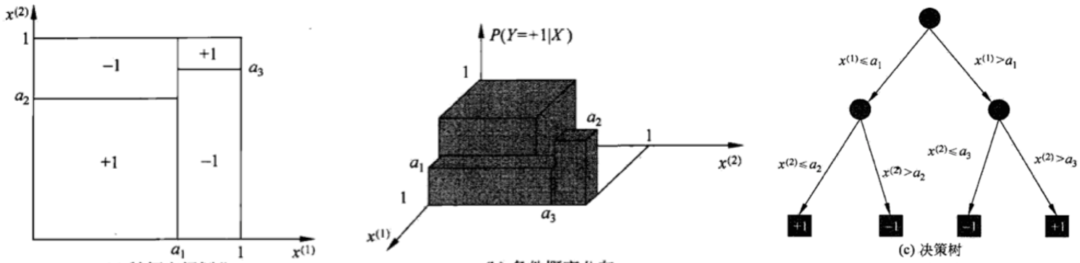


$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta^{\beta-1})$$
$$p(\theta|x_1,\cdots,x_n) = \frac{p(\theta,x_1,\cdots,x_n)}{p(x1,\cdots,x_n)}$$
$$= \frac{\pi(\theta)\cdot p(x_1|\,\theta),\cdots,p(x_n|\,\theta)}{\int p(\theta,x_1,\cdots,x_n)d\theta}$$
$$\propto \theta^{\alpha-1}\left(1-\theta^{\beta-1}\right)\prod \theta^{x_i}(1-\theta)^{1-x_i}$$
$$= \theta^{\sum x_i+\alpha-1}(1-\theta)^{n-\sum x_i+\beta-1}$$
$$\hat{\theta} = \frac{\sum x_i + \alpha - 1}{n+\alpha+\beta-2}$$

∴

$$MLE: \hat{\theta} = \frac{\sum x_i}{n}$$
$$BE: \hat{\theta} = \frac{\sum x_i + \alpha - 1}{n+\alpha+\beta-2} \overset{n\to\infty}{\longrightarrow} \frac{\sum x_i}{n}$$

5. Decision tree（if-then 規則）



(a) 特征空间划分　(b) 条件概率分布　(c) 决策树

图 5.2　决策树对应于条件概率分布

5.1 Entropy

S.t $P(X = x_i) = p_i, i = 1,2,\cdots,n$

Then X の Entropy：（もし $p_i = 0$ になれば $0\log 0 = 0$がある、普通 Base 2 或は e logarithm、Entropy 単位は Bit 或 Nat を呼ぶ）

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i$$

Entropy が X の分布だけで依頼する、Entropy おおきになるほど r.v の不確定性が大きにある

$$0 \le H(p) \le \log n$$

例えば：r.v={0,1}がある時、X の分布：

$$P(X = 1) = p, \quad P(X = 0) = 1 - p, \quad 0 \le p \le 1$$

Entropy：

$$H(p) = -p\log_2 p - (1-p)\log_2(1-p)$$

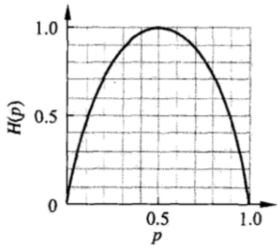


图 5.4　分布为贝努利分布时熵与概率的关系

5.1.1 Conditional entropy

$S.t\, r.v = (X, Y)$; Joint probability distribution:

$$P(X = x_i, Y = y_i) = p_{ij}, \quad i = 1,2,\cdots,n; \quad j = 1,2,\cdots,m$$

$$H(Y|X) = \sum_{i=1}^{n} p_i H(Y|X = x_i); \quad p_i = P(X = x_i); \quad i = 1,2,\cdots,n$$

もし Entropy と Conditional entropy がもらった確率の方法が MLE になる時 empirical entropy と empirical conditional entropy を呼ぶ。

5.1.2 Information gain

Input : Training data D & Feature A.
Output : Information gain g(D,A).

empirical entropy :

$$H(D) = -\sum_{k=1}^{k} \frac{|C_k|}{|D|}\log_2\frac{|C_k|}{|D|}$$

empirical conditional entropy :

$$H(D|A) = \sum_{i=1}^{n} \frac{|D_i|}{|D|} H(D_i)$$

$$= -\sum_{i=1}^{n}\frac{|D_i|}{|D|}\sum_{k=1}^{k}\frac{|D_{ik}|}{|D|}\log_2\frac{|D_{ik}|}{|D|}$$

Information gain :

　Training data D に対して Feature A の Information gain は$g(D,A)$を表示する。その中$H(D)-H(D|A)$のことが mutual information を呼ぶ。

$$g(D,A)=H(D)-H(D|A)$$

Information gain ratio:

　は Feature の類が多くになれば Information gain を影響する（Entropy=0 ; Information gain＝∞）。

$$g_R(D,A)=\frac{g(D,A)}{H_A(D)}$$

5.2 ID3 アルゴリズム

方法： Root node の中に可能性がある Feature の Information gain を計算して、可能性が大きな Feature をこの Root node の Feature になる

5.2.1 C4.5 アルゴリズム

方法：ID3 アルゴリズムの中に Information gain を Information gain ratio に替わること。

5.3 Pruning

　S.t T の leaf node 数= $|T|$。tはTの leaf node の一つ、N_tはこの leaf node の中にサンプル数、k類のサンプル数はN_{tk}コ。$H_t(T)$は t の leaf node の empirical entropy。

　Loss function : もし$C_\alpha(T)$は減られば Pruning をする

$$C_\alpha(T)=C(T)+\alpha|T|$$

$$H_t(T)=-\sum_{k}\frac{N_{tk}}{N_t}\log\frac{N_{tk}}{N_t}$$

$$C(T)=\sum_{t=1}^{|T|}N_tH_t(T)$$

5.4 CART（classification and regression tree）アルゴリズム

Breiman-1984

Classification 方法： C4.5 或 ID3 と違い所は CART が Binary tree と Gini index を利用する

Gini index：k=2(Binary tree)

$$Gini(p)=\sum_{k=1}^{k}p_k(1-p_k)=1-\sum_{k=1}^{k}p_k^2$$

Feature=A の D の Gini index

$$Gini(D,A)=\frac{|D_1|}{|D|}Gini(D_1)+\frac{|D_2|}{|D|}Gini(D_2)$$

X=p, Y=Loss

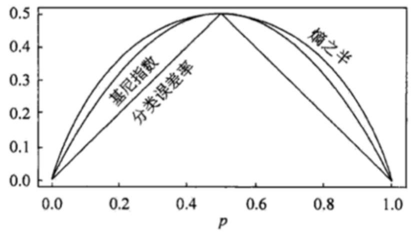


图 5.7 二类分类中基尼指数、熵之半和分类误差率的关系

Regression 方法
分散程度の度量方法が違う。Classification が Gini index を使う、
Regression の方が Squared difference を使う

6. Logistic regression & Maximum entropy
*** Logistic distribution**

Sigmoid curve : Center point:$\left(\mu,\frac{1}{2}\right)$; $F(-x+\mu)-\frac{1}{2}=-F(x-\mu)+\frac{1}{2}$

$$F(x)=P(X\le x)=\frac{1}{1+e^{-\frac{x-\mu}{\gamma}}}$$

$$f(x)=F'(x)=\frac{e^{-\frac{x-\mu}{\gamma}}}{\gamma\left(1+e^{-\frac{x-\mu}{\gamma}}\right)^2}$$

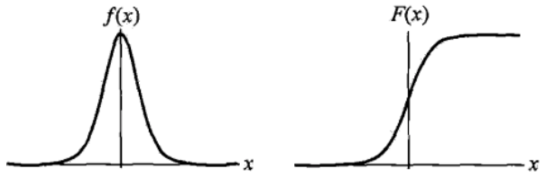


图 6.1 逻辑斯谛分布的密度函数与分布函数

6.1 Binomial logistic regression model
* Log-linear Model (Logistic regression & Conditional Random Field)
Log odds = logit function :

$$logit(p)=\log\frac{p}{1-p}$$

S.t : $y\in[0,1]$

$$P(Y=1|x)=\pi(x)$$

$$logit\big(\pi(x)\big)=\log\frac{\pi(x)}{1-\pi(x)},\quad \frac{\pi(x)}{1-\pi(x)}\in[0,+\infty],\log\frac{\pi(x)}{1-\pi(x)}\in[-\infty,+\infty]$$

$$\ln\frac{\pi(x)}{1-\pi(x)}=w\cdot x\Longrightarrow\pi(x)=\boldsymbol{P(Y=1|x)}=\frac{\mathbf{exp}(\boldsymbol{w\cdot x})}{\mathbf{1-exp}(\boldsymbol{w\cdot x})}$$

$$1-\pi(x)=\boldsymbol{P(Y=0|x)}=\frac{\mathbf{1}}{\mathbf{1-exp}(\boldsymbol{w\cdot x})}$$

MLE : **$\boldsymbol{w}$をもらい、$\boldsymbol{x}$と$\boldsymbol{y}$の$\boldsymbol{Joint\ density}$を最大限にする**
Probability distribution:

$$P_w(y|x) = \pi(x)^y[1-\pi(x)]^{1-y}$$

$$\max L(w) = \prod_{i=1}^{N} \pi(x)^{yi}[1-\pi(x)]^{1-yi}$$

$$\max \ln L(w) = \sum_{i=1}^{N} \{y_i \ln \pi(x) + (1-y_i)\ln[1-\pi(x)]\}$$

$$= \sum_{i=1}^{N} \left\{y_i \ln \frac{\pi(x_i)}{1-\pi(x_i)} + \ln[1-\pi(x)]\right\}$$

$$= \sum_{i=1}^{N} \{y_i(w \cdot x_i) - \ln[1+\exp(w \cdot x_i)]\}$$

SGD(max)

Multi-nominal logistic regression model

$$\ln \frac{P(Y=k|x)}{P(Y=K|x)} = w_i \cdot x\,; \qquad i = \#(K)$$

$$P(Y=k|x) = \frac{\exp(w_k \cdot x)}{1+\sum_{k=1}^{K-1}\exp(w_k \cdot x)}\,, \quad k = 1,2,\cdots,K-1$$

$$P(Y=K|x) = \frac{1}{1+\sum_{k=1}^{K-1}\exp(w_k \cdot x)}$$

6.2 Maximum entropy model

6.2.1Lagrangian: Equality Constraint

Constrained optimization →$\boxed{\text{Lagrangian}}$→ no restrained optimization

$$\min f(x)\,,\ s.t\ g(x) = 0\,; \quad \rightarrow \quad \min f(x) + \lambda g(x)$$

例えば：

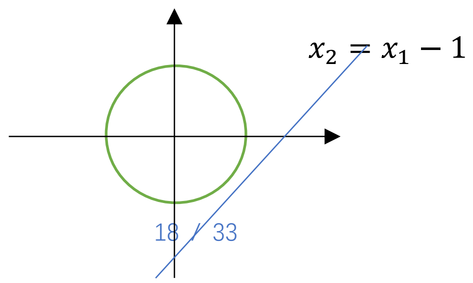
$$\min x_1^2 + x_2^2\,,\ s.t\ x_2 - x_1 = -1\,; \quad \rightarrow \quad \min f(x) + \lambda(x_2 - x_1 + 1)$$

$$\begin{pmatrix} \dfrac{\partial[f(x)+\lambda(x_2-x_1+1)]}{\partial x_1} \\ \dfrac{\partial[f(x)+\lambda(x_2-x_1+1)]}{\partial x_2} \\ \dfrac{\partial[f(x)+\lambda(x_2-x_1+1)]}{\partial \lambda} \end{pmatrix} = \begin{pmatrix} 2x_1 - \lambda \\ 2x_2 - \lambda \\ x_2 - x_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}\,; \qquad \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ \frac{-1}{2} \end{pmatrix}$$

Intuition of Equality Constraint

$$\min f(x) + \lambda g(x)$$

例：$\min x_1^2 + x_2^2\,,\ s.t\ x_2 - x_1 = -1$



$$\nabla_x f(x) || \nabla_x g(x)$$
$$\nabla_x f(x) = \pm\lambda \nabla_x g(x)$$
$$\nabla_x f(x) \pm \lambda \nabla_x g(x) = 0$$
$$\nabla_x\big(f(x)+\lambda g(x)\big) = 0\,; \quad \leftrightarrow \quad \frac{\partial\big(f(x)+\lambda g(x)\big)}{\partial x} 0$$
$$\nabla_x\big(f(x)+\lambda g(x)\big) = 0\,; \quad \leftrightarrow \quad \frac{\partial\big(f(x)+\lambda g(x)\big)}{\partial \lambda} 0\,; \quad \Longrightarrow \quad g(x) = 0$$

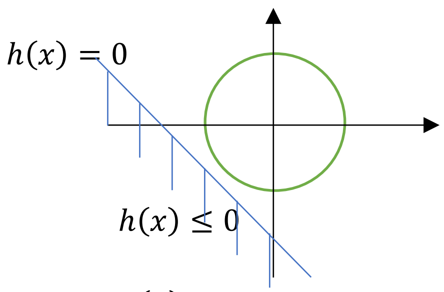
Multiple Equalities :

$s.t\ g_i(x) = 0\,, \quad i = 1,2,\cdots,R$

$$\min f(x) + \sum_{i=1}^{R} \lambda_i g_i(x)$$

Inequality Constraint : $\lambda \geq 0$

$s.t\ h(x) \leq 0\,;\ \min f(x)$



$$\begin{cases} \lambda = 0, & h(x) \leq 0 \\ h(x) = 0, & \lambda > 0 \end{cases}\;; \quad \Longrightarrow \quad \lambda h(x) = 0$$

$$s.t\ h(x) \leq 0\,;\ \min f(x) \quad \Longrightarrow \quad s.t\ \lambda h(x) = 0\ ;\min f(x) + \lambda h(x)$$

Appendix C :

$\min\limits_{x \in R^n} f(x)$

$s.t.\, c_i(x) \leq 0\,, i = 1,2,\cdots,k$

$\quad h_i(x) = 0, j = 1,2,\cdots,k$

Lagrangian :

$$L(x,\alpha,\beta) = f(x) + \sum_{i=1}^{k} \alpha_i c_i(x) \sum_{j=1}^{l} \beta_j h_j(x)$$

$$P \Leftrightarrow \min_x \max_{\alpha,\beta} L(x,\alpha,\beta) = \min_x \begin{cases} f(x), & c_i(x) \leq 0, h_i(x) = 0 \\ \infty, & other \end{cases}$$

Dualization:

$$Proof\!: \min_x \max_{\alpha,\beta} L(x,\alpha,\beta) \Leftrightarrow \max_{\alpha,\beta}\ \min_x L(x,\alpha,\beta)$$

$s.t.\ \alpha_i \geq 0$

$\alpha^* = opt(\alpha)$

$\beta^* = opt(\beta)$

$$d^* = opt\left(\max_{\alpha,\beta}\min_{x} L(x,\alpha,\beta)\right)$$

$$p^* = opt\left(\min_{x}\max_{\alpha,\beta} L(x,\alpha,\beta)\right)$$

p^* *'s Lowerbound* :

$$\max_{\alpha,\beta}\min_{x} L(x,\alpha,\beta) \le \max_{\alpha,\beta}\min_{x\in feasible\ zone} L(x,\alpha,\beta)$$

$$\le \min_{x\in feasible\ zone} f(x)$$

$$\therefore d^* \le p^*$$

$$if\ :\ s.t.\ \begin{cases}Convex\ optimization\\ Slater\end{cases} \Longrightarrow d^* = p^*(Strong\ duality) = L(x^*,\alpha^*,\beta^*)$$

$$\text{KKT condition:}\begin{cases}\nabla_x L(x^*,\alpha^*,\beta^*) = 0\\ c_i(x^*) \le 0 \ \rightarrow (Condition\ of\ the\ original\ problem)\\ h_j(x^*) = 0 \rightarrow (Condition\ of\ the\ original\ problem)\\ \alpha_i^* \ge 0 \rightarrow (Condition\ of\ the\ Dual\ problem)\\ \alpha_i^* c_i(x^*) = 0 \rightarrow (Complementary\ relaxation\ condition)\end{cases}$$

6.2.2 Maximum entropy mode

$$\max_{p\in C} H(P) = -\sum_{x,y} \tilde{P}(x)\, P(y|x) \log P(y|x)$$

$$s.t.\ \ E_p(f_i) = E_{\tilde{P}}(f_i),\ \ i = 1,2,\cdots,n$$

$$\sum_{y} P(y|x) = 1$$

MLE of Maximum entropy mode:

$$L(P,w) = -H(P) + w_0\left(1 - \sum_{y} P(y|x)\right) + \sum_{i=1}^{n} w_i\big(E_{\tilde{P}}(f_i) - E_P(f_i)\big)$$

$$= \sum_{x,y} \tilde{P}(x)P(y|x)\log P(y|x) + w_0\left(1 - \sum_{y} P(y|x)\right)$$

$$+ \sum_{i=1}^{n} w_i\left(\sum_{x,y} \tilde{P}(x,y) f_i(x,y) - \sum_{x,y} \tilde{P}(x)P(y|x) f_i(x,y)\right)$$

$$P_w(y|x) = \frac{1}{Z_w(x)}\exp\left(\sum_{i=1}^{n} w_i f_i(x,y)\right)$$

$$Z_w(x) = \sum_{y}\exp\left(\sum_{i=1}^{n} w_i f_i(x,y)\right)$$

$$w = argmax\ L_{\tilde{P}}(P_w) = \log \prod P(y|x)^{P(x,y)}$$

Proof of improved iterative scaling (IIS):
Log likelihood:

①

$$L(w) = \sum_{x,y}\left[\tilde{P}(x,y)\sum_{i=1}^{n} w_i f_i(x,y)\right] - \sum_{x}[\tilde{P}(x)\ln Z_w(x)]$$

$$L(w+\delta) - L(w) = \sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n}\delta_i f_i - \sum_{x}\tilde{P}(x)\ln\frac{Z_{w+\delta}(x)}{Z_w(x)}$$

②

$$\because -\ln x \geq 1-x \qquad \therefore \sum_{x}\tilde{P}(x) - \ln\frac{Z_{w+\delta}(x)}{Z_w(x)} \geq \sum_{x}\tilde{P}(x)\left[1-\frac{Z_{w+\delta}(x)}{Z_w(x)}\right]$$

$$\geq 1 - \sum_{x}\tilde{P}(x)\frac{Z_{w+\delta}(x)}{Z_w(x)}$$

②→③

$$\begin{aligned}
\frac{Z_{w+\delta}(x)}{Z_w(x)} &= \frac{1}{Z_w(x)}\sum_{y}\exp\left(\sum_{i=1}^{n}(w_i+\delta_i)f_i(x,y)\right) \\
&= \frac{1}{Z_w(x)}\sum_{y}\exp\left(\sum_{i=1}^{n}w_i f_i(x,y) + \sum_{i=1}^{n}\delta_i f_i(x,y)\right) \\
&= \frac{1}{Z_w(x)}\sum_{y}\left(\exp\sum_{i=1}^{n}w_i f_i(x,y)\right)\left(\exp\sum_{i=1}^{n}\delta_i f_i(x,y)\right) \\
&= \sum_{y}\left[\frac{1}{Z_w(x)}\left(\exp\sum_{i=1}^{n}w_i f_i(x,y)\right)\right]\left(\exp\sum_{i=1}^{n}\delta_i f_i(x,y)\right) \\
&= \sum_{y}\left[P(y|x)\left(\exp\sum_{i=1}^{n}\delta_i f_i(x,y)\right)\right]
\end{aligned}$$

①+③→ ④ : *Lowerbound*

$$L(w+\delta) - L(w) = \sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n}\delta_i f_i - \sum_{x}\tilde{P}(x)\ln\frac{Z_{w+\delta}(x)}{Z_w(x)}$$

$$\geq \sum_{x,y} \tilde{P}(x,y) \sum_{i=1}^{n} \delta_i f_i$$

$$- \sum_{x} \tilde{P}(x) \sum_{y} \left[P(y|x) \left(\exp \sum_{i=1}^{n} \delta_i f_i(x,y) \right) \right]$$

⑤$Jensen\ inequation$

$s\ t.\ \rho\ , a_i\ ; \sum a_i = 1$

$$\rho\left(\sum a_i x_i\right) \leq \sum a_i \rho(x_i)$$

$$\exp \sum_{i=1}^{n} \delta_i f_i(x,y) = \exp\left(\sum_{i} \frac{f_i(x,y)}{f^{\#}(x,y)} f^{\#} \delta_i \right)$$

$$\leq \sum_{i} \frac{f_i(x,y)}{f^{\#}(x,y)} \exp(f^{\#} f_i(x,y))$$

④→⑥

$L(w+\delta) - L(w)$

$$\geq \sum_{x,y} \tilde{P}(x,y) \sum_{i=1}^{n} \delta_i f_i - \sum_{x} \tilde{P}(x) \sum_{y} \left[P(y|x) \left(\exp \sum_{i=1}^{n} \delta_i f_i(x,y) \right) \right]$$

$$\geq \sum_{x,y} \tilde{P}(x,y) \sum_{i=1}^{n} \delta_i f_i - \sum_{x} \tilde{P}(x) \sum_{y} \left[P(y|x) \left(\sum_{i} \frac{f_i(x,y)}{f^{\#}(x,y)} \exp(f^{\#} f_i(x,y)) \right) \right]$$

7. SVM

1) hard interval: linearly separable, Sensitive to support vectors

$s.t.\, y_i(w \cdot x_i + b) \geq 1$

$$\max_{\mathrm{w,b}} \min_{\mathrm{i}} \frac{y_i(w \cdot x_i + b)}{||w||} = \max_{\mathrm{w,b}} \frac{1}{||w||} \min_{\mathrm{i}} y_i(w \cdot x_i + b)$$

$$= \max_{\mathrm{w,b}} \frac{1}{||w||}$$

$$= \min_{\mathrm{w,b}} ||w||$$

$$= \min_{\mathrm{w,b}} \frac{1}{2} ||w||^2$$

Dual problem

$$s.t. \sum_{i=1}^{N} \alpha_{\mathrm{i}} y_i = 0$$

$$\alpha_{\mathrm{i}} \geq 0$$

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_{\mathrm{i}} \alpha_{\mathrm{j}} y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{N} \alpha_{\mathrm{i}}$$

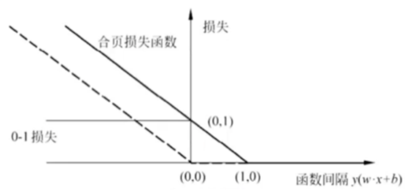
$$KKT \begin{cases} w^* = \sum_{i=1}^{N} \alpha_i^1 y_i x_i \\ b^* = y_j - \sum_{i=1}^{N} \alpha_{\mathrm{i}}^* y_i (x_i \cdot x_j) \end{cases} \quad \rightarrow \quad w^* \cdot x + b^* = 0$$

2) Soft interval

$$s.t. y_i (w \cdot x_i + b) \geq 1 - \xi$$

$$\xi_i \geq 0$$

$$\min_{\mathrm{w,b,\xi}} \frac{1}{2} \left||w|\right|^2 + C \sum_{i=1}^{N} \xi_i \Rightarrow \sum_{i=1}^{N} [1 - y_i (w \cdot x_i + b)]_+ + \lambda \left||w|\right|^2$$



Dual problem

$$s.t. \sum_{i=1}^{N} \alpha_{\mathrm{i}} y_i = 0$$

$$C \geq \alpha_{\mathrm{i}} \geq 0$$

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_{\mathrm{i}} \alpha_{\mathrm{j}} y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{N} \alpha_{\mathrm{i}}$$

$$KKT \begin{cases} w^* = \sum_{i=1}^{N} \alpha_i^* y_i x_i \cdot x \\ b^* = y_j - \sum_{i=1}^{N} \alpha_i^* y_i (x_i \cdot x_j) \end{cases} \quad \rightarrow \quad w^* \cdot x + b^* = 0$$

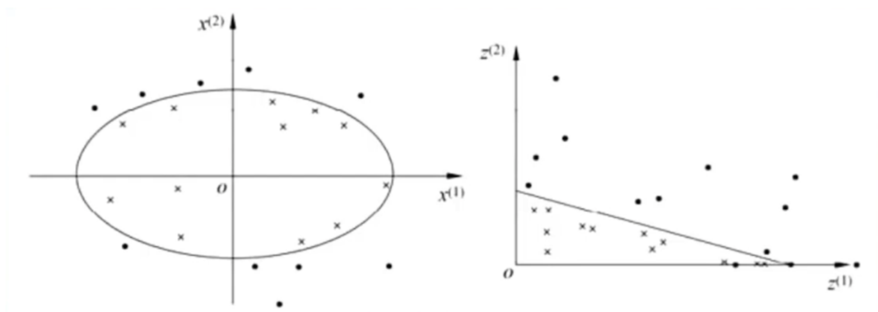


图7.7 非线性分类问题与核技巧示例

Dual problem

$$s.t. \sum_{i=1}^{N} \alpha_i y_i = 0$$

$$C \geq \alpha_i \geq 0$$

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i \cdot x_j) - \sum_{i=1}^{N} \alpha_i$$

$$f(x) = sign\left(\sum_{i=1}^{N} \alpha_i^* y_j K(x \cdot x_i) + b^*\right)$$

Proof : Dual problem

$1 - \xi_i - y_i(w \cdot x_i + b) \leq 0, -y_i \leq 0 \longrightarrow w, b, y \in Convex\ set$

$$L(w, b, \xi, \alpha, \mu) = \frac{1}{2}||w||^2 + C\sum \xi_i + \sum \alpha_i[1 - \xi_i - y_i(w \cdot x_i + b)] + \sum \mu_i - \xi_i$$

$$= \frac{1}{2}||w||^2 + C\sum \xi_i - \sum \alpha_i[y_i(w \cdot x_i + b)] - 1 + \xi_i] - \sum \mu_i \xi_i$$

① $\nabla_x L(x^*, \alpha^*, \beta^*) = 0$

$$\nabla_w L = w - \sum \alpha_i y_i x_i = 0 \Longrightarrow w = \sum \alpha_i y_i x_i$$

$$\nabla_b L = -\sum \alpha_i y_i = 0$$

$$\nabla_{\xi_i} L = C - \alpha_i - \mu_i = 0$$

$$L(w, b, \xi, \alpha, \mu) = \frac{1}{2}\left(\sum \alpha_i y_i x_i\right) \cdot \left(\sum \alpha_i y_i x_i\right) + C\sum \xi_i - \left(\sum \alpha_i y_i x_i\right)\left(\sum \alpha_i y_i x_i\right)$$
$$- b\sum \alpha_i y_i + \sum \alpha_i - \sum \alpha_i \xi_i - \sum \mu_i \xi_i$$
$$= \frac{1}{2}\left(\sum \alpha_i y_i x_i\right) \cdot \left(\sum \alpha_i y_i x_i\right) + C\sum \xi_i - \left(\sum \alpha_i y_i x_i\right)\left(\sum \alpha_i y_i x_i\right) + \sum \alpha_i$$
$$- \sum (\alpha_i - \mu_i)\xi_i$$

$$= \frac{1}{2}\left(\sum \alpha_i y_i x_i\right) \cdot \left(\sum \alpha_i y_i x_i\right) + C \sum \xi_i - \left(\sum \alpha_i y_i x_i\right)\left(\sum \alpha_i y_i x_i\right) + \sum \alpha_i - \sum C \xi_i$$

$$= \frac{1}{2}\left(\sum \alpha_i y_i x_i\right) \cdot \left(\sum \alpha_i y_i x_i\right) - \left(\sum \alpha_i y_i x_i\right)\left(\sum \alpha_i y_i x_i\right) + \sum \alpha_i$$

$$= -\frac{1}{2}\left(\sum \alpha_i y_i x_i\right) \cdot \left(\sum \alpha_i y_i x_i\right) + \sum \alpha_i$$

$$= \min_{\alpha} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i \cdot x_j) - \sum_{i=1}^{N} \alpha_i$$

② $c_i(x^*) \leq 0 \rightarrow (Condition\ of\ the\ original\ problem)$

$$y_i(w \cdot x_i + b) \geq 1 - \xi$$
$$\xi_i \geq 0$$

③ $h_j(x^*) = 0 \rightarrow (Condition\ of\ the\ original\ problem)$

④ $\alpha_i^* \geq 0 \rightarrow (Condition\ of\ the\ Dual\ problem)$

$$\mu_i \geq 0$$
$$\alpha_i \geq 0$$

⑤ $\alpha_i^* c_i(x^*) = 0 \rightarrow (Complementary\ relaxation\ condition)$

$$\alpha_i[y_i(w \cdot x_i + b)] - 1 + \xi_i] = 0$$
$$\mu_i \xi_i = 0$$

The uniqueness of hyperplane existence.

Proof existence: $w_1^*, b_2^* = w_1^*, b_2^*$

$$w = \frac{w_1^* + w_2^*}{2}\ ; \qquad b = \frac{b_1^* + b_2^*}{2}$$

$$y_i\left(\frac{w_1^* + w_2^*}{2} \cdot x_i + \frac{b_1^* + b_2^*}{2}\right) - 1 \geq 0?$$

$$y_i\left(\frac{w_1^* + w_2^*}{2} \cdot x_i + \frac{b_1^* + b_2^*}{2}\right) - 1 = \frac{1}{2}[y_i(w_1^* \cdot x_i + b_1^*) + y_i(w_2^* \cdot x_i + b_2^*) - 2]$$

$$\frac{1}{2}[y_i(w_1^* \cdot x_i + b_1^*) - 1 + y_i(w_2^* \cdot x_i + b_2^*) - 1] \geq 0$$

$s.t.\ \left||w_1^*|\right| = \left||w_2^*|\right| = C$

$$\left||w_1^*|\right| \leq \left||w|\right|$$

$$C \leq \left||w|\right| = \left\|\frac{1}{2}w_1^* + \frac{1}{2}w_2^*\right\| \leq \frac{1}{2}\left||w_1^*|\right| + \frac{1}{2}\left||w_2^*|\right| = C$$

$$\left\|\frac{1}{2}w_1^* + \frac{1}{2}w_2^*\right\| = \frac{1}{2}\left||w_1^*|\right| + \frac{1}{2}\left||w_2^*|\right|$$

8. Boosting

Adaboost

$input: Data\ set: T = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n),\}$

$x_i \in X \in \mathbb{R}^n, \qquad y_i \in Y = \{-1, +1\}$

$output: Classifier: G(x)$

$$e_m = P(G_m(x_i) \neq y_i) = \sum_{i=1}^{\tilde{N}} w_{mi} I(G_m(x_i) \neq y_i)$$

G_mが大きいほどα_mは小さくなる（$e_m < 0.5$；$\alpha_m > 0$）

$$\alpha_m = \frac{1}{2}\log\frac{1-e_m}{e_m}$$

重み更新；$y_i G_m(x_i) \in \{-1,1\}$, 予測違う時$-1$になり、重みが大きになる。

$$w_{m+1} = \frac{w_{mi}}{Z_m}\exp\left(-\alpha_m y_i G_m(x_i)\right)$$

正規化

$$Z_m \sum_{i=1}^{N} w_{mi} \exp\left(-\alpha_m y_i G_m(x_i)\right)$$

$G(x) Classifier: \alpha_m$は$G_m(x)$ $Classifier$の重み。

$$f(x) = \sum_{m=1}^{M} \alpha_m G_m(x)$$

$$G(x) = sign\left(f(x)\right) = sign\left(\sum_{m=1}^{M} \alpha_m G_m(x)\right)$$

解釈：

モデル＝足し算モデル

$\beta_m: primary\ function$

$$f(x) = \sum_{m=1}^{M} \beta_m b(x; \gamma_m)$$

戦略=Exponential function Loss

$y, x \in \{-1,1\}$

$$L\left(y, f(x)\right) = \exp\left[-yf(x)\right]$$

アルゴリズム= Forward step algorithm

モデル＝Boost tree(Classification tree)

$$f_m(x) \sum_{m=1}^{M} T(x; \theta_m)$$

戦略＝Square loss

$$L(y, f(x)) = \frac{1}{2}(y - f(x))^2$$

アルゴリズム＝Forward step algorithm

$$f_m(x) = f_{m-1}(x) + T(x; \theta_m)$$

$$\hat{\theta}_m = argmin \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + T(x_i; \theta_m))$$

9 Expectation Maximization (EM)

$s.t.\ Observed\ data\ Y$

$Hidden\ data\ Z$

$Model\ parameters\ \theta$

$input$: $P(Y, Z|\theta)$

$P(Z|Y, \theta)$

$output$: θ

Then:

$$P(Z_i|y_i, \theta_i^0) \to E(Z) \to \log \prod_{i=1}^{N} (Y_i, Z_i|\theta) \to \theta^{i+1} = \underset{\theta}{argmax} \log \prod_{i=1}^{N} (y_i, E(Z)|\theta)$$

1） $Initialization: \theta^0$

2） E

$$Q(\theta, \theta^i) = E_Z[\log P(Y, Z|\theta)|Y, \theta^i]$$

$$= \sum_Z \log P(Y, Z|\theta) P(Z|Y, \theta^i)$$

3） M

$$\theta^{i+1} = \underset{\theta}{argmax}\ Q(\theta, \theta^i)$$

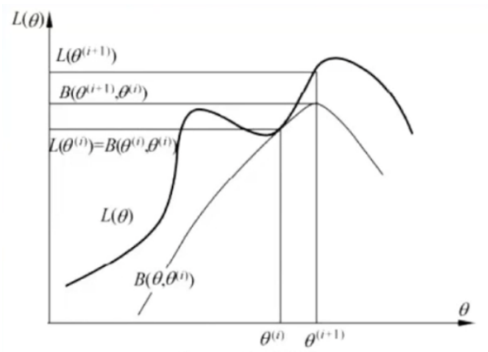


图9.1 EM算法的解释

Gaussian mixture model (GMM):

$y \sim N(\mu_i, \sigma_i^2)$

$\alpha \geq 0$; $\phi(y|\theta_k)$: $Gauss\ density$; $\theta_k = (\mu_k, \sigma_k^2)$

$$P(y|\theta) = \sum_Z p(y, z|\theta)$$

$$= \sum_{Z} p(Z|\theta)p(y|Z,\theta)$$

$$= \sum_{k=1}^{K} \alpha_k \phi(y|\theta_k)$$

$$\phi(y|\theta_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right)$$

$$Hidden\ v.r:\ \gamma_1(\gamma_{11}, \gamma_{12}, \dots \gamma_{1k}) = (1,0,0,\dots 0) \to \alpha_1$$
$$\vdots$$
$$= (0,0,0,\dots 1) \to \alpha_k$$

$$p(\gamma_1, y_1|\theta) = p(\gamma_1|\theta) \cdot p(y_1|\gamma_1,\theta)$$

$$= \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1k} \\ \vdots & \ddots & \vdots \\ \gamma_{k1} & \cdots & \gamma_{kk} \end{pmatrix} (\alpha_1, \alpha_2, \dots, \alpha_k) \cdot \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1k} \\ \vdots & \ddots & \vdots \\ \gamma_{k1} & \cdots & \gamma_{kk} \end{pmatrix} (y_1, y_2, \dots, y_k)$$

$$= \prod_{k=1}^{K} [\alpha_k \phi(y_1|\theta_k)]^{\gamma_{1k}}$$

$$p(\gamma, y|\theta) = \prod_{j=1}^{N} \prod_{k=1}^{K} \left[\alpha_k \phi\left(y_j\middle|\theta_k\right)\right]^{\gamma_{jk}}$$

$$= \prod_{k=1}^{K} \alpha_k^{\sum_{j=1}^{N} \gamma_{jk}} \prod_{j=1}^{N} \left[\phi\left(y_j\middle|\theta_k\right)\right]^{\gamma_{jk}}$$

$$= \prod_{k=1}^{K} \alpha_k^{\sum_{j=1}^{N} \gamma_{jk}} \prod_{j=1}^{N} \left[\frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{\left(y_j-\mu_k\right)^2}{2\sigma_k^2}\right)\right]^{\gamma_{jk}}$$

$$\ln p(\gamma, y|\theta) = \sum_{k=1}^{K} \left\{\sum_{j=1}^{N} \gamma_{jk} \ln \alpha_k + \sum_{j=1}^{N} \gamma_{jk} \left[\ln \frac{1}{\sqrt{2\pi}} - \ln \sigma_k - \frac{\left(y_j-\mu_k\right)^2}{2\sigma_k^2}\right]\right\}$$

E:

$$E\left[\sum_{j=1}^{N} \gamma_{jk}\right] = \sum_{j=1}^{N} E[\gamma_{jk}]$$

$$E\left[\gamma_{jk}|\theta^i, y\right] = P\left(\gamma_{jk} = 1\middle|\theta^i, y\right)$$

$$= \frac{P\left(\gamma_{jk} = 1, y_j\middle|\theta^i\right)}{P\left(y_j\middle|\theta^i\right)}$$

$$= \frac{P\left(\gamma_{jk} = 1, y_j \middle| \theta^i\right)}{\sum_{k=1}^{K} P\left(\gamma_{jk} = 1, y_j \middle| \theta^i\right)}$$

$$= \frac{P\left(y_j \middle| \gamma_{jk} = 1, \theta^i\right) P\left(\gamma_{jk} = 1 \middle| \theta^i\right)}{\sum_{k=1}^{K} P\left(y_j \middle| \gamma_{jk} = 1, \theta^i\right) P\left(\gamma_{jk} = 1 \middle| \theta^i\right)}$$

$$= \frac{\alpha_k \phi(y_i|\theta_k^i)}{\sum_K \alpha_k \phi(y_i|\theta_k^i)}$$

$$Q\left(\theta, \theta^i\right) = \mathrm{E}[\ln p(\gamma, y|\theta)]$$

$$= \sum_{k=1}^{K} \left\{ \sum\nolimits_{j=1}^{N} \gamma_{jk} \ln \alpha_k + \sum_{j=1}^{N} \gamma_{jk} \left[\ln \frac{1}{\sqrt{2\pi}} - \ln \sigma_k - \frac{\left(y_j - \mu_k\right)^2}{2\sigma_k^2} \right] \right\}$$

$$= \sum_{k=1}^{K} \left\{ \sum\nolimits_{j=1}^{N} \frac{\alpha_k \phi(y_i|\theta_k^i)}{\sum_K \alpha_k \phi(y_i|\theta_k^i)} \ln \alpha_k \right.$$

$$\left. + \frac{\alpha_k \phi(y_i|\theta_k^i)}{\sum_K \alpha_k \phi(y_i|\theta_k^i)} \sum_{j=1}^{N} \left[\ln \frac{1}{\sqrt{2\pi}} - \ln \sigma_k - \frac{\left(y_j - \mu_k\right)^2}{2\sigma_k^2} \right] \right\}$$

M: P9.30-9.32

$$\begin{cases} \dfrac{\partial Q\left(\theta, \theta^i\right)}{\partial \mu_k} = 0 \\ \dfrac{\partial Q\left(\theta, \theta^i\right)}{\partial \sigma_k^2} = 0 \\ \begin{cases} \dfrac{\partial Q\left(\theta, \theta^i\right)}{\partial} = 0 \\ \displaystyle\sum_K \alpha_k = 1 \end{cases} \end{cases} \quad \Longrightarrow \quad \begin{matrix} \mu_k^{i+1} \\ \sigma_k^{2(i+1)} \\ \alpha_k^{i+1} \end{matrix}$$

11 Hidden Markov model (HMM)

Probabilistic Graphical Model(GPM):Directed graph

$$\begin{matrix} i_1 & \to i_2 & \to \cdots & \to & i_t & \to & i_{t+1} & \to \cdots \to & i_{T-1} & \to i_T \\ \downarrow & \downarrow & & & \downarrow & & \downarrow & & \downarrow & \downarrow \\ o_1 & \to o_2 & \to \cdots & \to & o_t & \to & o_{t+1} & \to \cdots \to & o_{T-1} & \to i_T \end{matrix}$$

State transition matrix

$i_t \in \{q_1, q_2, \dots, q_N\}$

$o_t \in \{v_1, v_2, \dots, v_N\}$

$I = \{i_1, i_2, \dots, i_t\}$

$O = \{o_1, o_2, \dots, o_T\}$

$$A_{N*N} = \begin{pmatrix} a_{11} = P(i_1 = q_1|i_2 = q_1) & \cdots & a_{1N} = P(i_1 = q_1|i_2 = q_N) \\ \vdots & \ddots & \vdots \\ a_{N1} = P(i_1 = q_n|i_2 = q_1) & \cdots & a_{NN} = P(i_1 = q_N|i_2 = q_N) \end{pmatrix}$$

Observation probability matrix

$$B_{N*M} = \begin{pmatrix} b_1(1) = P(i_1 = q_1|o_1 = v_1) & \cdots & b_1(M) = P(i_1 = q_1|o_1 = v_M) \\ \vdots & \ddots & \vdots \\ b_N(1) = P(i_1 = q_N|o_1 = v_1) & \cdots & b_N(M) = P(i_1 = q_N|o_1 = v_M) \end{pmatrix}$$

Initial state probability vector

$$\pi_{N*1} = \begin{pmatrix} \pi_1 = P(i_1 = q_1) \\ \pi_2 = P(i_1 = q_2) \\ \vdots \\ \pi_N = P(i_1 = q_N) \end{pmatrix}$$

HMM parameters

$$\lambda = \begin{cases} \pi \ : \ N{+}{=}\ (N + 1 * Restrictions) \\ A \ : \ N{+}{=}\ (N * N + N * Restrictions) \\ B \ : \ N{+}{=}\ (N * M + N * Restrictions) \end{cases}$$

Homogeneous Markovian

Observational independence hypothesis

1) **Probability calculation problem:** $\boldsymbol{P(O|\lambda)}$

2) **Learning problem:** $\boldsymbol{\underset{\lambda}{argmax}\, P(O|\lambda)}$

3) **Prediction problem:** $\boldsymbol{\underset{I}{argmax}\, P(I|O)}$

Direct calculation: $O(TN^T)$

$$P(O|\lambda) = \sum_I P(O|I,\lambda)P(I|\lambda)$$

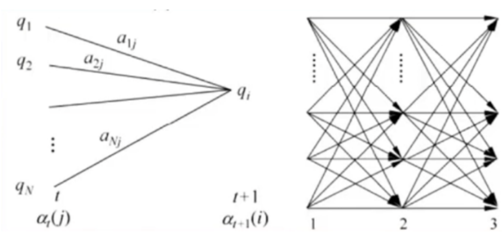
$$= \sum_{i_1,i_2,\ldots,i_T} \pi_{i_1} b_{i_1}(o_1) a_{i_1 i_2} b_{i_2}(o_2) \ldots a_{i_{T-1} i_T} b_{i_T}(o_T)$$

Forward algorithm: $O(TN^2)$

$\alpha_1(i) = \pi_i b_i(O_1), i = 1,2,\ldots,N$

$$\alpha_{i+1}(i) = \left[\sum_{j=1}^{N} \alpha_i(j) a_{ji}\right] b_i(o_{i+1}), i = 1,2,\ldots,N$$

$$P(O|\lambda) = \sum\nolimits_{i=1}^{N} \alpha_T(i)$$



Proof:

$$\alpha_t(i) = \sum_{i=1}^{N} P(o_1, o_2, \dots, o_T, i_T = q_i)$$
$$= \sum_{i=1}^{N} \alpha_T(i)$$

$$\alpha_t(j) = P\big(o_1, o_2, \dots, o_t, i_t = q_j\big)$$

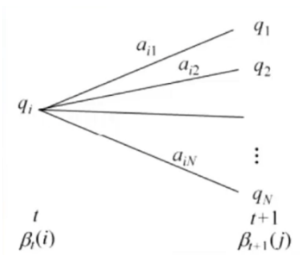
$$\begin{aligned}\alpha_{t+1}(i) &= P(o_1, o_2, \dots, o_{t+1}, i_{t+1} = q_i) \\ &= \sum_{j=1}^{N} P\big(o_1, o_2, \dots, o_{t+1}, i_{t+1} = q_i, i_t = q_j\big) \\ &= \sum_{j=1}^{N} P\big(o_1, o_2, \dots, o_t, i_t = q_j\big) P(o_{t+1}|i_{t+1} = q_i) P\big(i_{t+1} = q_i\big|i_t = q_j\big) \\ &= \left[\sum_{j=1}^{N} \alpha_i(j) a_{ji}\right] b_i(o_{i+1})\end{aligned}$$

Backward Algorithm: $O(TN^2)$

$\beta_T(i) = 1; i = 1,2, \dots, N$

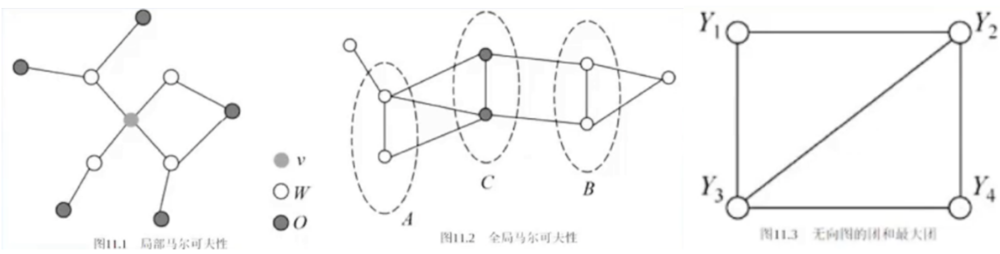
$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{i+1}) \beta_{i+1}(j)\, ; i = 1,2, \dots, N; t = T-1, T-2, \dots, 1$$

$$P(O|\lambda) = \sum\nolimits_{i=1}^{N} \pi_i b_i(o_1) \beta_1(i)$$



11 Conditional random field （CRF）

Undirected graph model



Hammersley-Clifford:

Joint distribution of undirected graphs

$$P(Y)=\frac{1}{Z}\prod_{C}\Psi_C(Y_C)$$

$$Z=\sum_{Y}\prod_{C}\Psi_C(Y_C)$$

$$\Psi_C(Y_C)=\exp\{-E(Y_C)\}$$

11.3:

$$P(y_1,y_2,y_3,y_4)=\frac{1}{Z}\psi_1(y_1,y_2,y_3)\cdot\psi_2(y_2,y_3,y_4)$$

CRF:

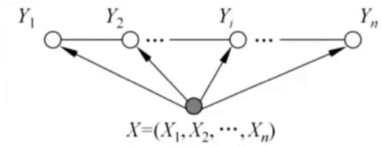


图11.4 线性链条件随机场

$$P(Y_v|X,Y_w,w\neq v)=P(Y_v|X,Y_w,w\sim v)$$

$$Z(x)=\sum_{y}\exp\left(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\right)$$

$$P(y|x)=\frac{1}{Z(x)}\exp\left(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\right)$$

$$=\frac{1}{Z(x)}\exp\sum_{k=1}^{K}w_k f_k(y,x)\Longrightarrow Simplified\ form$$

$$=P_w(y|x)=\frac{1}{Z_w(x)}\prod_{i=1}^{n+1}M_i(y_{i-1},y_i|x)\Longrightarrow\ Matrix\ form$$

Proof: Matrix from

$$\exp\left(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\right)$$

$$=\exp\left\{\sum_{i}\left[\sum_{k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{l}\mu_l s_l(y_i,x,i)\right]\right\}$$

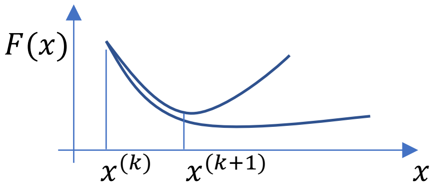
$$=\prod_{i}\exp\left[\sum_{k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{l}\mu_l s_l(y_i,x,i)\right]$$

$$=\prod_{i=1}^{n+1}M_i(y_{i-1},y_i|x)$$

$$M_i(x)=\begin{bmatrix}y_{i-1}=1;y_i=1 & \cdots & y_{i-1}=1;y_i=m\\ \vdots & \ddots & \vdots\\ y_{i-1}=m;y_i=1 & \cdots & y_{i-1}=m;y_i=m\end{bmatrix}_{m*m}$$

Appendix B

Newton Method:



$$f(x)=f\left(x^{(k)}\right)+\nabla f\left(x^{(k)}\right)^T\left(x-x^{(k)}\right)+\frac{1}{2}\left(x-x^{(k)}\right)^T H\left(x^{(k)}\right)\left(x-x^{(k)}\right)$$

$$\nabla f(x)=\nabla f\left(x^{(k)}\right)+H_k\left(x-x^{(k)}\right)=0$$

$$\Longrightarrow x^{(k+1)}=x^{(k)}-H_k^{-1}\nabla f\left(x^{(k)}\right)$$

Quasi-Newton Method:

$\nabla f(x)=\nabla f\left(x^{(k)}\right)+H_k\left(x-x^{(k)}\right)$

$H_k\left(x-x^{(k)}\right)=\nabla f(x)-\nabla f\left(x^{(k)}\right)$

$x-x^{(k)}=H_k^{-1}\left(\nabla f(x)-\nabla f\left(x^{(k)}\right)\right)$

でもH_k^{-1}は$x^{(k+1)}$を計算しかない

$x-x^{(k)}=H_{k+1}^{-1}\left(\nabla f(x)-\nabla f\left(x^{(k)}\right)\right)$

DFP: $G_k\rightarrow H_k^{-1}$

$$G_{k+1}=G_k+avv^T+buu^T$$

$$G_{k+1}=G_k+\frac{\left(x-x^{(k)}\right)\left(x-x^{(k)}\right)^T}{\left(x-x^{(k)}\right)^T\left(\nabla f(x)-\nabla f\left(x^{(k)}\right)\right)}-\frac{G_k\left(\nabla f(x)-\nabla f\left(x^{(k)}\right)\right)\left(\nabla f(x)-\nabla f\left(x^{(k)}\right)\right)^T G_k}{\left(\nabla f(x)-\nabla f\left(x^{(k)}\right)\right)^T G_k\left(\nabla f(x)-\nabla f\left(x^{(k)}\right)\right)}$$

BFGS: $B_k\rightarrow H_k$

P223-SM