情報意味論 (第6回)
ベイズ推論とナイーブベイズ

慶應義塾大学理工学部
櫻井 彰人

目次

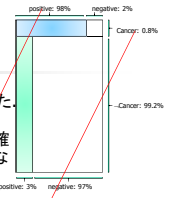
- Bayes 定理
- MAP と ML
- Bayes 最適分類器, Gibbs アルゴリズム
- クラスの推定か確率の推定か
- Naïve Bayes

Bayes の定理

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}$$

$$P(A,B) = P(A \mid B)P(B)$$
$$= P(B \mid A)\,P(A)$$

例 (Mitchell Chap. 6.2)



ある患者がガンの検査を受けたところ結果が陽性であった。この患者には、本当に病変があるのだろうか？
なお、当該検査は、本当に病変があるときに陽性となる確率は 98% を誇る. また、病変がないときに正しく陰性となる確率は 97% である.

さらに, 全人口に対するこのガンをもつ率は .008 である.

P(cancer) = .008 P(¬cancer) = .992
P(+ | cancer) = .98 P(– | cancer) = .02
P(+ | ¬cancer) = .03 P(– | ¬cancer) = .97

P(+) = P(+ | c'r) P(c'r) + P(+ | ¬c'r) P(¬c'r)=.0376

$$\text{P(cancer} \mid +) = \frac{P(+ \mid cancer)\,P(cancer)}{P(+)} = \mathbf{.209}$$

例 (Mitchell Exercise 6.1)

2回目の検査(2回は統計的に独立とする)を受け, その結果も陽性であったとしよう. ガンである事後確率はどうなるであろうか?

P(cancer) = .008 P(¬cancer) = .992
P(+ | cancer) = .98 P(– | cancer) = .02
P(+ | ¬cancer) = .03 P(– | ¬cancer) = .97

$P(+_1 +_2) = P(+_1 +_2 \mid c'r)\,P(c'r) + P(+_1 +_2 \mid \neg c'r)\,P(\neg c'r) = .00858$

$$P(\text{cancer} \mid +_1 +_2) = \frac{P(+_1 +_2 \mid cancer)\,P(cancer)}{P(+_1 +_2)} = \mathbf{.896}$$

良く使う公式

乗法の公式(実は、条件付確率の定義！):

$P(A \wedge B) = P(A \mid B)\,P(B) = P(B \mid A)\,P(A)$

参考: 和事象に対しては

$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

全確率の公式:

$$P(B) = \sum_{i=1}^{n} P(B, A_i) = \sum_{i=1}^{n} P(B \mid A_i)\,P(A_i)$$

仮説選択に関して教えてくれること

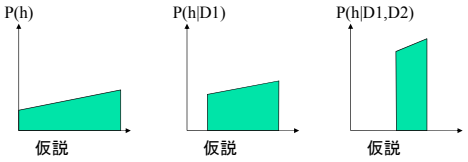
$$P(h \mid D) = \frac{P(D \mid h)\,P(h)}{P(D)}$$

P(h) = 仮説 h の事前確率
P(D) = 訓練データ D の生起確率
P(h|D) = D が与えられたときの h の事後確率
P(D|h) = h が与えられたときの D の生起確率

データ D を生成したらしい仮説 h を選択することができる！

注: 条件付確率は因果関係（もしあれば）を反映するわけではない
注: "仮説の生起確率" を考えることができるのだろうか？

ノイズがないときの事後確率の進展



目次

- Bayes 定理
- MAP と ML
- Bayes 最適分類器, Gibbs アルゴリズム
- クラスの推定か確率の推定か
- Naïve Bayes

MAP推定

$$P(h \mid D) = \frac{P(D \mid h)\,P(h)}{P(D)}$$

データが所与のとき、必要とするのは、最もありうべき仮説であろう.
事後確率最大仮説（Maximum a posteriori hypothesis） h_{MAP}:

$$\begin{aligned} h_{MAP} &= \arg\max_{h \in H} P(h \mid D) \\ &= \arg\max_{h \in H} \frac{P(D \mid h)\,P(h)}{P(D)} \\ &= \arg\max_{h \in H} P(D \mid h)\,P(h) \end{aligned}$$

ML推定

全ての i, j について $P(h_i) = P(h_j)$ と仮定すれば, より簡単化でき, 最尤Maximum Likelihood (ML) 仮説 を選ぶことになる

$$h_{MAP} = \arg\max_{h \in H} P(D \mid h)\,P(h)$$
$$h_{ML} = \arg\max_{h \in H} P(D \mid h)$$

ML推定の一つの解釈

- 現実世界では、事前確率分布は、未知か、計算不能か、存在しないと思われる
  - 例えば、文書における単語の生起頻度の事前分布はあるのだろうか？　年齢、社会的背景、人口分布で大きく異なりうる
- 事前確率分布が存在しないとしたら、尤度最大化は自然な考え

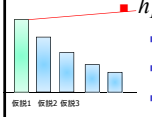
尤度最大化は、各仮説の生起確率がすべて等しいとした場合と等価である。
つまり仮説の事前確率分布が一様であるとの仮定と等価である。
妥当か？

目次

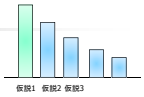
- Bayes 定理
- MAP と ML
- Bayes 最適分類器, Gibbs アルゴリズム
- クラスの推定か確率の推定か
- Naïve Bayes

未知事例の最もありうる分類

- これまで、事例 D のもとでの最もありうる仮説を求めてきた（例：h_{MAP}）。
- 未知事例の最もありうる（最も確率が高い）分類結果はどうなるのであろうか？
  - $h_{MAP}(x)$ は最もありうる分類ではない！
    - 次の例で、x のもっともありうる類別は？
    - 3仮説: $P(h_1|D)$=0.4, $P(h_2|D)$=0.3, $P(h_3|D)$=0.3
    - 新事例: $h_1(x)$=+, $h_2(x)$=–, $h_3(x)$= –



Bayes 最適な分類器



$$\arg\max_{c_j \in \{+,-\}} \sum_{h_i \in H} P(c_j \mid h_i) P(h_i \mid D)$$

注: Bayes 最適な分類器は H に含まれるとは限らない

注: 論文にはうまくいくと報告されているのだが、試してみるとMAPやMLと変わらない場合がある。どのような場合にそうなるか、興味のあるところである

注: 実行可能か？　見るからに時間がかかりそう

例 (Mitchell Chap. 6.7)

$P(h_1 \mid D) = .4$	$P(- \mid h_1) = 0$	$P(+ \mid h_1) = 1$
$P(h_2 \mid D) = .3$	$P(- \mid h_2) = 1$	$P(+ \mid h_2) = 0$
$P(h_3 \mid D) = .3$	$P(- \mid h_3) = 1$	$P(+ \mid h_3) = 0$

それゆえ:

$$\sum_{h_i \in H} P(+ \mid h_i) P(h_i \mid D) = .4$$

$$\sum_{h_i \in H} P(- \mid h_i) P(h_i \mid D) = .6$$

そして:

$$\arg\max_{c_j \in \{+,-\}} \sum_{h_i \in H} P(c_j \mid h_i) P(h_i \mid D) = -$$

Bayes最適な分類器

- パラメータθをもつ確率分布 P(X;θ) からn個のデータ D={x_1,…,x_n } が観測されたとする。D に基づき、次のデータ y がなんであるかを推定したい。
- 方法1: パラメータθを推定し、P(X;θ)に基づき推定する
  - MLE（最尤推定）　$\theta_{MLE} = \arg\max P(D|\theta)$
  - MAP（事後確率最大化）　$\theta_{MAP} = \arg\max P(D|\theta)P(\theta)$
  - 期待値（posterior mean）

$$\hat{\theta} = \int \theta P(\theta|D) d\theta = \int \theta P(D|\theta)P(\theta)/P(D) d\theta$$

- 方法2: パラメータθを推定しないで求める

$$P(Y,\theta|D) = P(Y,D|\theta)P(\theta)/P(D)$$

$$\Rightarrow \quad P(Y|D) = \int P(Y,D|\theta)P(\theta)/P(D) d\theta$$

ちょっと戻って、ゆっくり進んでみよう

ベイズ学習の基本コンセプト

- ベイズ的見方では、不確実さは計測可能である。たとえ、サンプル数が多くなくても。
  - 例えば、ある初出場の高校が甲子園で優勝する確率は？
    - 頻度主義では求めることができない
  - ある特定のコインを投げた時に、表が出る確率は？
  - いずれでも、多くのデータがない場合、まず、事前知識を信じることになる
- しかし、データが集まるにつれ、データをより一層信じるようになる（もっともらしく思うようになる）
- データが全て揃えば、事前知識は全く考慮しなくなる

18

ベイズ推論例

- 確率変数 x の平均 μ を学習したいとしよう。分散 σ^2 は既知だがデータはまだないとする。
- $P(\mu \mid D,\sigma^2) = P(D \mid \mu,\sigma^2)P(\mu)/P(D)$
  $\propto P(D \mid \mu,\sigma^2)P(\mu)$
- ベイジアンであれば、事前分布と尤度とをパラメータ付の関数・分布で表現しようと考える（分布としては、例えば、正規（Gaussian）, 多項, 一様 等）
- 正規分布を仮に考えよう
- 事前分布が正規布なので、尤度との積を作り、事後確率分布（尤度）を求めた結果もまたパラメータ付の分布になるが、これも正規分布であるとうれしい

共役分布

- $P(\mu \mid D, \sigma_0^2) = P(D \mid \mu)P(\mu)/P(D) \propto P(D \mid \mu)P(\mu)$
- もし、事後分布（事前分布と尤度関数との積）が事前分布と同じ形の関数であれば、その事前分布はその尤度関数に対する共役事前分布と呼ぶ
- 分散が既知の正規分布（ガウス分布）を事前分布とし、尤度関数が正規分布の場合、事後分布は正規分布となる。
- 尤度関数が多項分布関数の場合、ディリクレ（Dirichlet）分布が共役事前分布となる。

離散分布の共役分布

Discrete distributions [edit]

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters	Interpretation of hyperparameters[note 1]	Posterior predictive[note 2]
Bernoulli	p (probability)	Beta	α, β	$\alpha + \sum_{i=1}^{n} x_i, \beta + n - \sum_{i=1}^{n} x_i$	$\alpha - 1$ successes, $\beta - 1$ failures[note 1]	$p(\tilde{x}=1) = \frac{\alpha'}{\alpha'+\beta'}$
Binomial	p (probability)	Beta	α, β	$\alpha + \sum_{i=1}^{n} x_i, \beta + \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} x_i$	$\alpha - 1$ successes, $\beta - 1$ failures[note 1]	$\mathrm{BetaBin}(\tilde{x} \mid \alpha', \beta')$ (beta-binomial)
Negative Binomial with known failure number r	p (probability)	Beta	α, β	$\alpha + \sum_{i=1}^{n} x_i, \beta + rn$	$\frac{\alpha}{\beta - 1}$ total successes, $\beta - 1$ failures[note 1] (i.e. $\frac{\alpha}{\beta - 1} r$ experiments, assuming r stays fixed)	
Poisson	λ (rate)	Gamma	k, θ	$k + \sum_{i=1}^{n} x_i, \frac{\theta}{n\theta + 1}$	k total occurrences in $1/\theta$ intervals	$\mathrm{NB}(\tilde{x} \mid k', \frac{\theta'}{1+\theta'})$ (negative binomial)
Poisson	λ (rate)	Gamma	α, β[note 3]	$\alpha + \sum_{i=1}^{n} x_i, \beta + n$	α total occurrences in β intervals	$\mathrm{NB}(\tilde{x} \mid \alpha', \frac{1}{1+\beta'})$ (negative binomial)
Categorical	$\boldsymbol{p}$ (probability vector), k (number of categories, i.e. size of $\boldsymbol{p}$)	Dirichlet	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha} + (c_1, \ldots, c_k)$, where c_i is the number of observations in category i	$p(\tilde{x} = i) = \frac{\alpha_i'}{\sum \alpha_i'} = \frac{\alpha_i + c_i}{\sum \alpha_i + n}$	
Multinomial	$\boldsymbol{p}$ (probability vector), k (number of categories, i.e. size of $\boldsymbol{p}$)	Dirichlet	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha} + \sum_{i=1}^{n} \mathbf{x}_i$	$\alpha_i - 1$ occurrences of category i[note 1]	$\mathrm{DirMult}(\tilde{\mathbf{x}} \mid \boldsymbol{\alpha}')$ (Dirichlet-multinomial)
Hypergeometric with known total population size N	M (number of target members)	Beta-binomial[4]	$n = N, \alpha, \beta$	$\alpha + \sum_{i=1}^{n} x_i, \beta + \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} x_i$	$\alpha - 1$ successes, $\beta - 1$ failures[note 1]	
Geometric	p_0 (probability)	Beta	α, β	$\alpha + n, \beta + \sum_{i=1}^{n} x_i$	$\alpha - 1$ experiments, $\beta - 1$ total failures[note 1]	

Wikipediaより

連続分布の共役分布（1）

Continuous distributions [edit]

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters	Interpretation of hyperparameters	Posterior predictive[note 4]
Normal with known variance σ^2	μ (mean)	Normal	μ_0, σ_0^2	$\left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^{n} x_i}{\sigma^2}\right) \Big/ \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right), \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}$	mean was estimated from observations with total precision (sum of all individual precisions) $1/\sigma_0^2$ and with sample mean μ_0	$\mathcal{N}(\tilde{x} \mid \mu_0', \sigma_0'^2 + \sigma^2)$[6]
Normal with known precision τ	μ (mean)	Normal	μ_0, τ_0	$\left(\tau_0\mu_0 + \tau\sum_{i=1}^{n} x_i\right) \Big/ (\tau_0 + n\tau), \tau_0 + n\tau$	mean was estimated from observations with total precision (sum of all individual precisions) τ_0 and with sample mean μ_0	$\mathcal{N}\left(\tilde{x} \mid \mu_0', \frac{1}{\tau_0'} + \frac{1}{\tau}\right)$[6]
Normal with known mean μ	σ^2 (variance)	Inverse gamma	α, β[note 5]	$\alpha + \frac{n}{2}, \beta + \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2}$	variance was estimated from 2α observations with sample variance β/α (i.e. with sum of squared deviations 2β); where deviations are from known mean μ)	$t_{2\alpha}(\tilde{x} \mid \mu, \sigma^2 = \beta'/\alpha')$[6]
Normal with known mean μ	σ^2 (variance)	Scaled inverse chi-squared	ν, σ_0^2	$\nu + n, \frac{\nu\sigma_0^2 + \sum_{i=1}^{n}(x_i - \mu)^2}{\nu + n}$	variance was estimated from ν observations with sample variance σ_0^2	$t_{\nu'}(\tilde{x} \mid \mu, \sigma_0'^2)$[6]
Normal with known mean μ	τ (precision)	Gamma	α, β[note 5]	$\alpha + \frac{n}{2}, \beta + \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2}$	precision was estimated from 2α observations with sample variance β/α (i.e. with sum of squared deviations 2β); where deviations are from known mean μ)	$t_{2\alpha'}(\tilde{x} \mid \mu, \sigma^2 = \beta'/\alpha')$[6]
Normal[note 6]	μ and σ^2 Assuming exchangeability	Normal-inverse gamma	$\mu_0, \nu, \alpha, \beta$	$\frac{\nu\mu_0 + n\bar{x}}{\nu + n}, \nu + n, \alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{n\nu}{\nu + n}\frac{(\bar{x} - \mu_0)^2}{2}$ • $\bar{x}$ is the sample mean	mean was estimated from ν observations with sample mean μ_0; variance was estimated from 2α observations with sample mean μ_0 and sum of squared deviations 2β	$t_{2\alpha'}\left(\tilde{x} \mid \mu', \frac{\beta'(\nu'+1)}{\alpha'\nu'}\right)$[6]
Normal	μ and τ Assuming exchangeability	Normal-gamma	$\mu_0, \nu, \alpha, \beta$	$\frac{\nu\mu_0 + n\bar{x}}{\nu + n}, \nu + n, \alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{n\nu}{\nu + n}\frac{(\bar{x} - \mu_0)^2}{2}$ • $\bar{x}$ is the sample mean	mean was estimated from ν observations with sample mean μ_0, and precision was estimated from 2α observations with sample mean μ_0 and sum of squared deviations 2β	$t_{2\alpha'}\left(\tilde{x} \mid \mu', \frac{\beta'(\nu'+1)}{\alpha'\nu'}\right)$[6]

Wikipediaより

連続分布の共役分布（2）

Multivariate normal with known covariance matrix $\boldsymbol{\Sigma}$	$\boldsymbol{\mu}$ (mean vector)	Multivariate normal	$\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0$	$(\boldsymbol{\Sigma}_0^{-1} + n\boldsymbol{\Sigma}^{-1})^{-1}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + n\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}),$ $(\boldsymbol{\Sigma}_0^{-1} + n\boldsymbol{\Sigma}^{-1})^{-1}$ • $\bar{\mathbf{x}}$ is the sample mean	mean was estimated from observations with total precision (sum of all individual precisions) $\boldsymbol{\Sigma}_0^{-1}$ and with sample mean $\boldsymbol{\mu}_0$	$\mathcal{N}(\tilde{\mathbf{x}} \mid \boldsymbol{\mu}_0', \boldsymbol{\Sigma}_0' + \boldsymbol{\Sigma})$[6]
Multivariate normal with known precision matrix $\boldsymbol{\Lambda}$	$\boldsymbol{\mu}$ (mean vector)	Multivariate normal	$\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0$	$(\boldsymbol{\Lambda}_0 + n\boldsymbol{\Lambda})^{-1}(\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 + n\boldsymbol{\Lambda}\bar{\mathbf{x}}), (\boldsymbol{\Lambda}_0 + n\boldsymbol{\Lambda})$ • $\bar{\mathbf{x}}$ is the sample mean	mean was estimated from observations with total precision (sum of all individual precisions) $\boldsymbol{\Lambda}$ and with sample mean $\boldsymbol{\mu}_0$	$\mathcal{N}\left(\tilde{\mathbf{x}} \mid \boldsymbol{\mu}_0', (\boldsymbol{\Lambda}_0^{-1} + \boldsymbol{\Lambda}^{-1})^{-1}\right)$[6]
Multivariate normal with known mean $\boldsymbol{\mu}$	$\boldsymbol{\Sigma}$ (covariance matrix)	Inverse-Wishart	$\nu, \boldsymbol{\Psi}$	$n + \nu, \boldsymbol{\Psi} + \sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T$	covariance matrix was estimated from ν observations with sum of pairwise deviation products $\boldsymbol{\Psi}$	$t_{\nu'-p+1}\left(\tilde{\mathbf{x}} \mid \boldsymbol{\mu}, \frac{1}{\nu'-p+1}\boldsymbol{\Psi}'\right)$[6]
Multivariate normal with known mean $\boldsymbol{\mu}$	$\boldsymbol{\Lambda}$ (precision matrix)	Wishart	$\nu, \mathbf{V}$	$n + \nu, \left(\mathbf{V}^{-1} + \sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T\right)^{-1}$	covariance matrix was estimated from ν observations with sum of pairwise deviation products $\mathbf{V}^{-1}$	$t_{\nu'-p+1}\left(\tilde{\mathbf{x}} \mid \boldsymbol{\mu}, \frac{1}{\nu'-p+1}\mathbf{V}^{-1}\right)$[6]
Multivariate normal	$\boldsymbol{\mu}$ (mean vector) and $\boldsymbol{\Sigma}$ (covariance matrix)	normal-inverse-Wishart	$\boldsymbol{\mu}_0, \kappa_0, \nu_0, \boldsymbol{\Psi}$	$\frac{\kappa_0\boldsymbol{\mu}_0 + n\bar{\mathbf{x}}}{\kappa_0 + n}, \kappa_0 + n, \nu_0 + n,$ $\boldsymbol{\Psi} + \mathbf{C} + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^T$ • $\bar{\mathbf{x}}$ is the sample mean • $\mathbf{C} = \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$	mean was estimated from n_0 observations with sample mean $\boldsymbol{\mu}_0$; covariance matrix was estimated from n_0 observations with sample mean $\boldsymbol{\mu}_0$ and with sum of pairwise deviation products $\boldsymbol{\Psi} = n_0\boldsymbol{\Sigma}_0$	$t_{n_0'-p+1}\left(\tilde{\mathbf{x}} \mid \boldsymbol{\mu}_0', \frac{n_0'+1}{n_0'(n_0'-p+1)}\boldsymbol{\Psi}'\right)$[6]
Multivariate normal	$\boldsymbol{\mu}$ (mean vector) and $\boldsymbol{\Lambda}$ (precision matrix)	normal-Wishart	$\boldsymbol{\mu}_0, \kappa_0, \nu_0, \mathbf{V}$	$\frac{\kappa_0\boldsymbol{\mu}_0 + n\bar{\mathbf{x}}}{\kappa_0 + n}, \kappa_0 + n, \nu_0 + n,$ $\left(\mathbf{V}^{-1} + \mathbf{C} + \frac{n_0 n}{n_0 + n}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^T\right)^{-1}$ • $\bar{\mathbf{x}}$ is the sample mean • $\mathbf{C} = \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$	mean was estimated from n_0 observations with sample mean $\boldsymbol{\mu}_0$; covariance matrix was estimated from n_0 observations with sample mean $\boldsymbol{\mu}_0$ and with sum of pairwise deviation products $\mathbf{V}^{-1}$	$t_{n_0'-p+1}\left(\tilde{\mathbf{x}} \mid \boldsymbol{\mu}_0', \frac{n_0'+1}{n_0'(n_0'-p+1)}\mathbf{V}^{-1}\right)$[6]

Wikipediaより

連続分布の共役分布（3）

Uniform	$U(0, \theta)$	Pareto	x_m, k	$\max\{x_1, \ldots, x_n, x_m\}, k + n$	k observations with maximum value x_m	
Pareto with known minimum x_m	α (shape)	Gamma	α, β	$\alpha + n, \beta + \sum_{i=1}^{n} \ln \frac{x_i}{x_m}$	α observations with sum β of the order of magnitude of each observation (i.e. the logarithm of the ratio of each observation to the minimum x_m)	
Weibull with known shape β	θ (scale)	Inverse gamma[4]	a, b	$a + n, b + \sum_{i=1}^{n} x_i^\beta$	a observations with sum b of the βth power of each observation	
Log-normal with known precision τ	μ (mean)	Normal[4]	μ_0, τ_0	$\left(\tau_0\mu_0 + \tau\sum_{i=1}^{n} \ln x_i\right) \Big/ (\tau_0 + n\tau), \tau_0 + n\tau$	"mean" was estimated from observations with total precision (sum of all individual precisions) τ_0 and with sample mean μ_0	
Log-normal with known mean μ	τ (precision)	Gamma[4]	α, β[note 5]	$\alpha + \frac{n}{2}, \beta + \frac{\sum_{i=1}^{n}(\ln x_i - \mu)^2}{2}$	precision was estimated from 2α observations with sample variance $\frac{\beta}{\alpha}$ (i.e. with sum of squared log deviations 2β) — i.e. deviations between the logs of the data points and the "mean")	
Exponential	λ (rate)	Gamma	α, β[note 5]	$\alpha + n, \beta + \sum_{i=1}^{n} x_i$	α observations that sum to β	$\mathrm{Lomax}(\tilde{x} \mid \beta', \alpha')$ (Lomax distribution)
Gamma with known shape α	β (rate)	Gamma	α_0, β_0	$\alpha_0 + n\alpha, \beta_0 + \sum_{i=1}^{n} x_i$	α_0 observations with sum β_0	$\mathrm{CG}(\tilde{\mathbf{x}} \mid \alpha, \alpha_0', \beta_0') = \beta'(\tilde{\mathbf{x}} \mid \alpha, \alpha_0', 1, \beta_0')$[note 7]
Inverse Gamma with known shape α	β (inverse scale)	Gamma	α_0, β_0	$\alpha_0 + n\alpha, \beta_0 + \sum_{i=1}^{n} \frac{1}{x_i}$	α_0 observations with sum β_0	
Gamma with known rate β	α (shape)	$\propto \frac{\alpha^{a-1}\beta^{bc}}{\Gamma(\alpha)^c}$	a, b, c	$a\prod_{i=1}^{n} x_i, b + n, c + n$	b or c observations (b for estimating α, c for estimating β) with product a	
Gamma[4]	α (shape), β (inverse scale)	$\propto \frac{p^{\alpha-1}e^{-\beta q}}{\Gamma(\alpha)^r \beta^{-s\alpha}}$	p, q, r, s	$p\prod_{i=1}^{n} x_i, q + \sum_{i=1}^{n} x_i, r + n, s + n$	α was estimated from r observations with product p; β was estimated from s observations with sum q	

Wikipediaより

ベイズ推論例

- 事前分布: $P(\mu) = N(\mu \mid \mu_0, \sigma_0^2)$
- 事後分布: $P(\mu \mid D) = N(\mu \mid \mu_N, \sigma_N^2)$

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

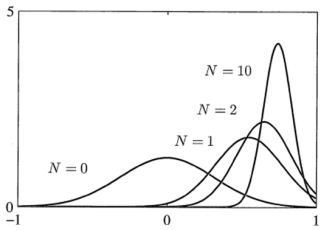
$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n \qquad \sigma^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu_{ML})^2$$

- 信じる度合が、事前分布からデータへと移る様子が分かる

25

ベイズ推論例

分散が既知とした時、ある正規分布の平均 μ をベイズ推論する場合の例図。曲線は、μ の事前分布（$N = 0$）（これも正規分布）, N を増加させていった場合の事後分布である。データ点は、平均値 0.8, 分散 0.1 の正規分布から生成した。また事前分布の平均値は0とした。事前分布と尤度関数では、分散は真の値とした



26

ベイズ推論例

- もしこの問題において、仮に、平均が既知で分散が未知とすると、共役事前分布は、逆ガンマ分布となる
  - もし、精度（分散の逆数）を用いるなら、共役事前分布はガンマ分布となる
- 平均と分散とが未知であれば（典型的な場合でしょう）共役事前分布は正規-逆ガンマ分布となる（正規分布と逆ガンマ分布の組合せ）。
- 通常は多変数であるので、これは、他変数正規-逆ガンマ分布となる。これは、逆ウィシャート分布として知られている

27

ベイズ推論例

- $P(\mu,\sigma^2 \mid D) = P(D \mid \mu,\sigma^2)P(\mu,\sigma^2)/P(D)$
  $\propto P(D \mid \mu,\sigma^2)\, P(\mu \mid \sigma^2)P(\sigma^2)$
- 事前分布: $P(\mu \mid \sigma^2) = N(\mu \mid \mu_0, \sigma^2/k_0)$,
  $P(\sigma^2) = \mathrm{IG}(\sigma^2 \mid r_0/2, s_0/2)$
  $N\_IG(\mu, \sigma^2 \mid \mu_0, k_0, r_0, s_0)$
- 事後分布: $P(\mu \mid \sigma^2, D) = N(\mu \mid \mu_N, \sigma^2/k_N)$,
  $P(\sigma^2) = \mathrm{IG}(\sigma^2 \mid r_N/2, s_N/2)$
  $N\_IG(\mu, \sigma^2 \mid \mu_N, k_N, r_N, s_N)$

$$\mu_N = \frac{k_0}{k_0 + N}\mu_0 + \frac{N}{k_0 + N}\mu_{ML} \qquad r_N = r_0 + N$$

$$k_N = k_0 + N \qquad s_N = r_0 + (N-1)$$

28

ベイズ推論

- ベイジアン（当然統計学者）ならば、ニューラルネットワークや決定木やSVMや最近傍法モデルを、尤度関数として用いるのは、気に食わない（と思う）
  - なぜ?

29

ベイズ推論

- ベイジアン（当然、統計学者）ならば、ニューラルネットワークや決定木やSVMや最近傍法モデルを、尤度関数として用いるのは、気にくわない（と思う）。
  - なぜか? – これらのモデルは標準的なパラメータ化した統計的な分布ではなく、当然ながら、事前分布との積を作って事後分布を作ることができない
  - ニューラルネットワークや決定木等に確率を出力させることはできる。しかし、真に確率分布になっているわけではない
    - Softmax を使う等.
- 確率分布が分かるのはいいことだが、しかし、最重要な点は、精度がよいことである
  - 勿論、正確な確率モデルが得られるならそれは大満足なのだが、もしそうでなければ、無理にということはない

30

ベイズ最適分類器

- 問うべきは、与えられた事例に対し、「何が最もありうる分類結果か」であって、与えられたデータセットに対し、「何が最もありうる仮説か」ではない
- ありうる全ての仮説を考え、与えられた事例に関し、予測する分類結果に対し重み付で投票することにする。重みは対応する仮説の事後確率とする。結果は、一般に、単独のMAP仮説よりよくなる

$$P(v_j \mid D, H) = \sum_{h_i \in H} P(v_j \mid h_i) P(h_i \mid D) = \sum_{h_i \in H} P(v_j \mid h_i) \frac{P(D \mid h_i) P(h_i)}{P(D)}$$

- ベイズ最適分類器:

$$\underset{v_j \in V}{\operatorname{argmax}} \sum_{h_i \in H} P(v_j \mid h_i) P(h_i \mid D) = \underset{v_j \in V}{\operatorname{argmax}} \sum_{h_i \in H} P(v_j \mid h_i) P(D \mid h_i) P(h_i)$$

31

ベイズ最適分類器

- データと仮説空間の事前分布が与えられている場合、同じ仮説空間を使う他の方法は、平均的には、ベイズ最適分類器を越えることはできない。
- 無限の、または有限でも非常に多数の、仮説からなる仮説空間を用いる場合、ベイズ最適分類器は、実用的ではない
- また、ベイズ最適分類器の正確さは、仮説に対する事前確率に関する知識（領域知識）の正確さ並みである。しかし、領域知識はしばしば得られない。
  - しかし、領域知識は、もし少しでもあるなら、それは大助かりである
  - 例えば, 自動的に過剰的が制御でき、validation set も early stopping も考える必要はない.
- しかし、事前知識が間違っている場合、ベイズ最適分類器は最適ではなくなる. 例えば, 一様な事前確率分布を想定すると, 事後確率が低い多くの仮説が、はるかに少ないが高い事後確率をもった仮説を凌駕してしまうことが発生する.
- とはいうものの, ベイズ最適分類器は理論的に重要な概念であり, ベイズ最適という概念に基づき、ただし、それを簡略化した、多くの実用的なアルゴリズムを生むきっかけともなっている

32

Gibbs 分類器 – 速度向上

1. 仮説を P(h|D) に従ってランダムに選ぶ
2. 新事例をこれに従い分類する

慶賀: もし仮説を事前分布 P(h) に従ってランダムに選ぶと,

$E[\text{error}_{\text{Gibbs}}] \le 2E[\text{error}_{\text{BayesOptimal}}]$

(詳細は “Mitchell Machine Learning Chap. 6.8”)

仮説の個数が多くて、ベイズ最適な分類器が計算できないときに有用

目次

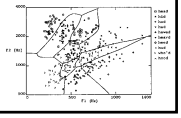
- Bayes 定理
- MAP と ML
- Bayes 最適分類器, Gibbs アルゴリズム
- クラスの推定か確率の推定か
- Naïve Bayes

学習目標値には2通り

- カテゴリ: 分類問題
  - （説明変数の）空間を部分空間に分割
    - 分割する境界を得る
  - カテゴリに数値を対応させる
- 連続値: 回帰問題
  - 離散値も連続値の一部と考える

連続関数を使う場合には、同じことになります。
(1) 境界を関数の零集合で表す
(2) 関数値が0以外のところが各カテゴリの空間

ただし、不連続関数になるので、回帰問題化には注意が必要
誤り数最小化とするのが妥当

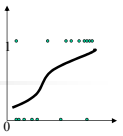


ところが

- カテゴリ値の場合（境界で0をとる連続関数を探すとき）
  - 関数値が0に近い＝境界に近い＝判断に迷い
  - 仮に、推定の確信度合いを、0から1の実数で表せば、カテゴリを表す部分空間で
    - 中ほど＝判断に自信＝値は1に近い、
    - 境界に近い＝判断に迷い＝値は0に近い

  とすると、回帰問題と考えられる。
    - 値は、カテゴリに振った番号

見方を変えて



- 確率とサンプル点の個数
  - ある点での（あるクラスに属するという）度合を、確定値ではなく、確率で考えることにすれば、その確率が、その付近にあるサンプル点の個数に現れると考えると考える
- 確信度合いとサンプル点の個数
  - ある点での（今いる部分空間にいるという）確信度合いはその付近にあるサンプル点の個数に比例すると考える

まとめると

目標値	カテゴリ値		連続値
考え方と方法	誤り数最小化（境界の推定）		目標値の回帰（誤差最小化）
	カテゴリ値を目標値とした回帰（誤差最小化）	出力値を丸めてカテゴリ値と解釈	
		出力値とカテゴリ値の違いを確率と解釈	
	入力点を確率分布のサンプルと考え、分布推定（密度推定）		

ニューラルネット学習方法2通り

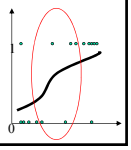
分類　　回帰

教えるのは分類クラス。しかし出力値を確率と解釈可能

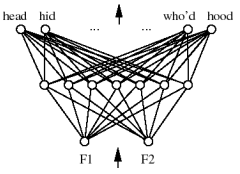
教えるのは分類クラス。出力値は丸めて分類クラスと解釈

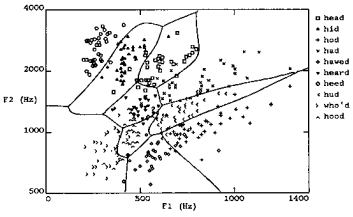
標準シグモイド関数を用いると、出力値は0と1の間

これが可能なのは、過学習しにくいからである（右図）
online 学習なら resampling していることとほぼ同じ



再掲: クラスと境界

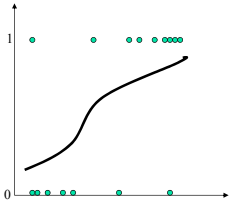




出力値は、実数であり、シグモイド関数を出力関数とすると、例えば[0,1]とすることができる。クラスごとに出力素子を用意すれば、そのクラスへの所属確率を出力しているとも解釈できる。
教師信号は、クラス種別でもよいが、その「クラスらしさ」ともできる。
後者の場合、確率の学習（回帰）とも考えることができる。

確率を予測するように学習する

クラス種別を教えたのに、確率を学習してしまうとは？

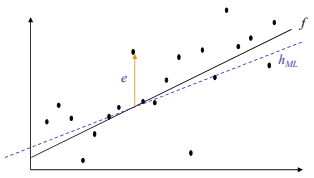


online学習で、頻度を学習する。
個数を数える。

ロジスティック回帰と比べてみよう！

実数値の学習（回帰分析）

ところで、回帰分析とは？



回帰分析の統計的解釈

学習事例: $<x_i, d_i>$ 但し

$d_i = f(x_i) + e_i$

e_i はノイズ = iid なる正規分布に従う確率変数で、平均=0 かつ分散は有限とする

iid=independent, identically distributed

ならば (予想): (次のスライドで証明) random variable

$$h_{ML} = \arg\min_{h \in H} \sum_{i=1}^{m} (d_i - h(x_i))^2$$

回帰分析の統計的解釈(証明)

$$\begin{aligned}
h_{ML} &= \arg\max_{h \in H} \ln p(D \mid h) \\
&= \arg\max_{h \in H} \ln \prod_{i=1}^{m} e^{-\frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2} \\
&= \arg\max_{h \in H} \sum_{i=1}^{m} -\frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2 \\
&= \arg\max_{h \in H} \sum_{i=1}^{m} -(d_i - h(x_i))^2 \\
&= \arg\min_{h \in H} \sum_{i=1}^{m} (d_i - h(x_i))^2
\end{aligned}$$

確率の予測には二乗誤差は不適

- 例: 生存確率を患者データから学習しよう

$$\begin{aligned}
h_{ML} &= \arg\max_{h \in H} \ln p(D \mid h) \\
&= \arg\max_{h \in H} \ln \prod_{i=1}^{m} P(d_i \mid h, x_i) P(x_i) \\
&= \arg\max_{h \in H} \sum_{i=1}^{m} \ln[P(d_i \mid h, x_i) P(x_i)] \\
&= \arg\max_{h \in H} \sum_{i=1}^{m} \ln\left(h(x_i)^{d_i} (1 - h(x_i))^{1-d_i} P(x_i)\right) \\
&= \arg\max_{h \in H} \sum_{i=1}^{m} d_i \ln h(x_i) + (1 - d_i) \ln(1 - h(x_i))
\end{aligned}$$

d_i は 0 or 1 (または所属確率)

注: cross entropy $H(p,q) = -\sum_x p(x) \log q(x) = H(p) + D_{KL}(p \| q)$

目次

- Bayes 定理
- MAP と ML
- Bayes 最適分類器, Gibbs アルゴリズム
- クラスの推定か確率の推定か
- Naïve Bayes

Naïve Bayes 分類器

- 単純だが(だから?)よく知られた分類方法
  - 単純な割には高精度
  - 単純なだけに、高速
- Bayes 定理 + 仮定 *条件付独立*
  - 実際には成り立たないことが多い仮定
  - それにも関わらず, 実際にはしばしばうまくいく
- 成功事例:
  - 文書分類
  - 診断

Naïve Bayes は Bayesian とは関係ない

Bayes 定理を使う場合の課題

- 変数 x の属性$<a_1,\ldots,a_n>$が与えられたとき, x が属するクラス c を最尤推定するには?

$$\begin{aligned}
c_{MAP} &= \arg\max_{c_j \in C} P(c_j \mid a_1, a_2, \ldots, a_n) \\
&= \arg\max_{c_j \in C} \frac{P(a_1, a_2, \ldots, a_n \mid c_j) P(c_j)}{P(a_1, a_2, \ldots, a_n)} \\
&= \arg\max_{c_j \in C} P(a_1, a_2, \ldots, a_n \mid c_j) P(c_j)
\end{aligned}$$

- 問題: 大量のデータが $P(a_1 \ldots a_n | c_j)$ を推定するのに必要. パラメータ数が膨大 ($\Pi|A_i|$) (2値属性の場合、属性数が n なら 2^n 個)だから

Naïve Bayes 分類器

- Naïve Bayes の仮定: 属性同士は、属するクラスが所与なら、独立
  - $P(a_1,\ldots,a_n|c_j) = P(a_1|c_j)\,P(a_2|c_j)\,\ldots\,P(a_n|c_j)$
  - *条件付独立性* (クラスが所与の時)
  - 推定すべきパラメータ数の削減:
    $\Pi|A_i|$ (=O(2^n)) $\rightarrow$ $\Sigma|A_i|$ (=O(n))
- この仮定のもと, c_{MAP} は

$$c_{NB} = \arg\max_{c_j \in C} P(c_j)\prod_i P(a_i \mid c_j)$$

Naïve Bayes: アルゴリズム

学習(事例集合)

事例がクラス c_j に属する確率

$\hat{P}(c_j) = P(c_j)$ の推定値

クラス c_j に属する事例の i 番目の属性の属性値が a_i である確率

$\hat{P}(a_i|c_j) = P(a_i|c_j)$ の推定値

分類(x)

$$c_{NB} = \arg\max_{c_j \in C} \hat{P}(c_j)\prod_i \hat{P}(a_i \mid c_j)$$

Naïve Bayes: 推定

- どうやって $P(c_j)$ と $P(a_i|c_j)$ を推定するか?
  - 統計学が教える標準的な方法
    - サンプルの頻度から確率を推定する
    - $P(c)$ の推定値は $\mathrm{count}(c)\,/\,N$
    - $P(A|B)$ の推定値は $\mathrm{count}(A \wedge B)\,/\,\mathrm{count}(B)$
  - 例: 100 事例. 内訳 70 + と 30 –
    - P(+)=0.7 かつ P(-)=0.3
    - 70 個の正例のなかに, 35 個で a_1=SUNNY
    - P(a_1=SUNNY|+)=0.5

例

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

$P(Y) = 9/14,$
$P(sunny\,|\,Y) = 2/9,$
$P(cool\,|\,Y) = 3/9,$
$P(high\,|\,Y) = 3/9,$
$P(strong\,|\,Y) = 3/9$

Naïve Bayes: 例

- 例の *PlayTennis* , と新事例
  <Outlk=sun, Temp=cool, Humid=high, Wind=strong>
- 計算したいのは:

$$c_{NB} = \arg\max_{c_j \in C} \hat{P}(c_j)\prod_i \hat{P}(a_i \mid c_j)$$

- $\hat{P}(Y)\hat{P}(sun\,|\,Y)\hat{P}(cool\,|\,Y)\hat{P}(high\,|\,Y)\hat{P}(strong\,|\,Y) = 0.005$
  $\hat{P}(N)\hat{P}(sun\,|\,N)\hat{P}(cool\,|\,N)\hat{P}(high\,|\,N)\hat{P}(strong\,|\,N) = 0.021$

$$\Rightarrow c_{NB} = No$$

Naïve Bayes: 条件付独立は必須か?

- もし仮定が成り立たなかったら?
  - i.e. if $P(a_1,\ldots,a_n|c_j) \neq P(a_1|c_j)\,P(a_2|c_j)\ldots P(a_n|c_j)$
- それでも、下記の(弱い)条件が成り立つ限り、予測値は Bayes 予測値と等価:

$$\arg\max_{c_j \in C} P(a_1 \mid c_j)P(a_2 \mid c_j)\ldots P(a_n \mid c_j)P(c_j)$$
$$= \arg\max_{c_j \in C} P(a_1, a_2, \ldots, a_n \mid c_j)P(c_j)$$

- しかし、予測時に求める *確率* は 0 や 1 に極めて近い非現実的な値になりうる

Naïve Bayes: ある問題

- もしも、あるクラス c_j で属性値 a_i が観測されなかったら？
  - 推定値 $P(a_i|c_j)=0$ なぜなら $\text{count}(a_i \wedge c_j) = 0$?
  - 影響は甚大: これが 0 だと積は 0!
- 解: Laplace correction を用いる
  - $\hat{P}(a_i \mid c_j) = \dfrac{n_c + mp}{n + m}$
  - n 訓練例数. 但し $c = c_j$
  - n_c 訓練例数. 但し $c = c_j$ かつ $a = a_j$
  - p 事前確率（の推定値）$\hat{P}(a_i|c_j)$ (通常は一様分布)
  - m “仮想” 事例数（しばしば、当該属性 a の属性値の個数を用いる）
    m=1 とする方法がある。その方が結果がよいことがある

補足: Laplace correction

- （度数から生起確率（パラメータ）を推定する時に）パラメータに事前分布を想定し、MAP推定を行う
- 事前分布として、ベータ分布
  $f(x;\alpha,\beta)=x^{\alpha-1}(1-x)^{\beta-1}/B(\alpha,\beta)$ を考える
- パラメータの posterior mean をとったものが Laplace correction である。Bernoulli 試行の場合、$\hat{\theta} = (n_0 + \alpha)/\big((n_0 + n_1) + \alpha + \beta\big)$ となる

補足: スムージング

- 確率モデルの推定において、訓練データに出現しない事象に対して微小な確率値を割り当てること。平滑化とも呼ばれる。
  http://www.jaist.ac.jp/project/NLP_Portal/doc/glossary/index.html
- 自然言語処理では、単語や文字 n 個の連なり（n-gram）の出現頻度をよく用いる。n が少し大きくなると、出現しない単語列・文字列が出てくる。それらを出現頻度0とするといろいろまずいことが起こる。そこで、様々なスムージング法が提案されてきた。
  - Laplaceスムージング（加算スムージング）
  - 線形補間法(Interpolation)
  - グッド・チューリング スムージング
  - カッツ スムージング
  - チャーチ・ゲイル スムージング
  - ウィトン・ベル スムージング
  - Kneser-Ney スムージング
  - ....
  - 階層的Pitman-Yor言語モデル

57

文書分類

- 文書分類とは:
  - 文書（メール、ニュース、webページ等々。それらの一段落ということも、また、一文ということも）を分類すること
  - 分りやすいのは、メールがスパムか否かの分類
  - ブログを、スブログか否かに分類する、とうい課題もある
  - ニュースが（ある人にとって）興味のあるものか否かを分類する、というのもある。さらに、
  - ある商品の評判を（良い評判も悪い評判も）集めるにも「分類」が必要。そして、良い評判と悪い評判とに分ける。
  - 信頼できる評価か信用できない評価かに分けるのも、文書分類
  - アンケート調査のうち、自由記述文の分類。
  - コールセンターでの、QAの分類
- Naïve Bayes が結構うまくいく
  - どうやって Naïve Bayes を用いるか?
  - ポイント: どう事例（すなわち、1文書）を表現するか? 属性は何か?

文書の表現方法

- Bag-of-words, すなわち、袋一杯の単語 or 袋詰めの単語
  - ある文書を、それぞれの単語が何回現れたかで表現する。
    - "Bag" で、もとの文書のどこにあったかを忘れることを表している。
    - また、単語の連なりも考えないことを表している。
    - 例えば、仮に、「慶應」「義塾」「大学」がそれぞれ単語なら、「慶應義塾大学」も「慶應大学義塾」も「義塾慶應大学」も同じと考えることになる。
  - 「何を単語とするか」が結構重要。文書ごとに変わってはいけない。
    - 英語であれば、dog と dogs といったような語形変化は無視した方がよい。
  - 文書分類に役立ちそうもない単語は考えない
    - 日本語で言えば、助詞（は、が、も、や、、、）がその代表。
    - 英語で言えば、前置詞がその代表
    - こういった、文法機能を持ち、単語単独では意味のない単語を機能語という
  - ノイズの可能性が高い単語は考えない。
    - 文書集合内で、出現頻度が極めて低い（一回等）もの

文書の表現方法（続）

- 表現自体が naïve Bayes 的
  - ベイズ推論とは直接には関係しないので、naïve Bayes ではないが、naïve な表現であることは間違いない。
  - しかし、naïve Bayes 的に、文書の出現確率を書くことができる。
  - 文書の属するクラスごとに、文書内にある特定の単語が出現する確率 $P(w_1|c_j), P(w_2|c_j), \ldots, P(w_n|c_j)$ が決まっているとすると、$w_1, w_2, \ldots, w_n$ が文書中に含まれる単語であるとき、そのような文書が出現する確率を次のように書く

    $P(\text{doc}|c_j)=P(w_1|c_j)^{\text{TF}(w1)} P(w_2|c_j)^{\text{TF}(w2)} \ldots P(w_n|c_j)^{\text{TF}(wn)}$

    ただし TF(w) は単語 w の doc における出現度数(term frequency)

出現確率をこう書けば naïve Bayes といえよう

Naïve Bayes による文書分類

- ある文書 doc につき

$$c_{NB} = \arg\max_{c_j \in C} P(c_j) \prod_{w_k \in Voc} P(w_k \mid c_j)^{TF(w_k, doc)}$$

ただし、$TF(w_k,\text{doc})$=doc中のw_kの出現度数、Voc は全単語（考えている全単語）集合とした

- 単語の出現確率については、Laplace correction が必須。そこで、下記の推定値を使用; ただし、n_j=クラスc_j中の全単語出現度数, $n_{k,j}$=クラスc_j中の単語w_k出現度数

$$P(w_k \mid c_j) = \frac{n_{k,j} + 1}{n_j + |Voc|}$$

Twenty News Groups (Joachims 1996)

- 各グループ1000の訓練文書
- 新規の文書を、もとのnewsgroupに割振る

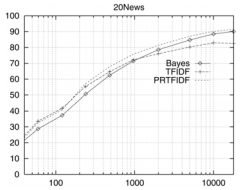
comp.graphics	misc.forsale
comp.os.ms-windows.misc	rec.autos
comp.sys.ibm.pc.hardware	rec.motorcycles
comp.sys.mac.hardware	rec.sport.baseball
comp.windows.x & rec.sport.hockey	rec.sport.hockey
alt.atheism	sci.space
soc.religion.christian	sci.crypt
talk.religion.misc	sci.electronics
talk.politics.mideast	sci.med
talk.politics.misc	
talk.politics.guns	

T. Joachims. *A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization.*
In Proceedings of the 14th International Conference on Machine Learning, Nashville, TN, 1997, pp.143--151.

Twenty News Groups (Joachims 1996)

- Naive Bayes: 89% 分類正解率
  - 頻出単語上位100 個 (the and of …) は除去
    - このように文法機能を担う単語や、文書を類別するのに有効でない単語を stop words として除去するのが普通
  - 頻度が3回に満たない単語は除去
  - 残った単語は、約 38,500 語

ただし、この正解率は高すぎ。20 Newsgroups の各投稿には、分類に非常に役立つ subject フィールドがある。今ではこれらは除去することになっているが、当時では、除去せずに、分類実験をした可能性がある。



精度対訓練データ数（1/3はテスト用にとりおいた）

20 Newsgroups: Rでは？

- データが多すぎて、Rのパッケージに含まれる naïve Bayes 分類器は使えない。
  - データ行列（さきほどのRプログラムでは、xy, xy, tt といった行列）が巨大になる（行数が文書数（約2,000）、列数が単語数（約40,000））。
  - しかし、非零要素は非常に少ないので、スパース行列表示を用いればよい。
  - それでもオーバーヘッドが大きい。
  - それなら自分でプログラムを書いてしまおう。
- なお、Weka にもスパース行列が表現できて、原理的には取り扱える。しかし、大きなメモリが必要で、しかも遅い。

20 Newsgroups: データ

- "20 Newsgroups" というサイトにあり
  - http://people.csail.mit.edu/jrennie/20Newsgroups/
- 前処理（単語の切り出し等）が終わって、単語の個数のデータに編集が終わったものを用いる。Matlab で使いやすい形になっている。
  - 20news-bydate-matlab.tgz
- このうち、train.data, train.label, test.data, test.label を用いる。
- プログラムは資料として web頁に掲載しておきます。
- 結果のうち、confusion matrix を次頁に示します。
- 正解率は、約78.2%です。

```
> cm
         correct
predicted   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
       1  237   3   3   0   0   0   1   0   4   2   0   2  10   3   7   2  12   7  47
       2    0 299  33   8   8  42   9   1   1   1   0   5  18   7   8   2   0   1   1   3
       3    0   7 208  15  10   8   4   0   0   0   0   1   0   1   0   1   0   0   0   0
       4    0  12  58 306  38  10  49   2   0   1   0   1  28   3   0   0   0   0   0   0
       5    0   7  11  21 275   2  21   0   0   1   0   2   8   0   0   1   1   0   0   0
       6    1  21  30   2   3 306   1   1   0   2   0   1   3   0   2   2   0   0   1   0
       7    0   1   0   4   4   1 227   5   1   3   1   1   1   1   0   0   2   0   0   0
       8    0   3   2   6   4   0  32 356  25   3   1   0   9   3   0   0   2   2   1   0
       9    0   1   2   0   1   2   5   4 353   1   0   0   2   0   1   0   1   1   0   1
       10   0   0   2   0   1   1   0   2   2 345   4   0   0   2   0   0   1   1   0   0
       11   1   0   1   1   0   0   1   0   0  16 381   0   0   0   1   0   0   1   0   0
       12   1  16  17   5   5  10   3   1   1   2   1 361  45   0   3   1   3   4   3   1
       13   1   4   1  23  16   0  11   4   1   2   0   3 260   3   4   0   0   0   0   0
       14   2   3   4   0   7   0   2   0   1   0   2   2   6 324   4   1   1   0   3   3
       15   3   6   4   1   2   3   3   2   0   0   1   0   3   3 333   0   2   0   7   5
       16  43   4   5   0   0   1   3   0   1   3   2   2   6  16   5 377   3   7   2  69
       17   3   0   0   0   3   1   1   5   4   1   0   7   0   3   1   2 324   3  95  19
       18   9   0   0   0   0   1   3   1   2   2   1   0   2   6   2   2   2 323   5   5
       19   7   2   9   0   6   2   6   9   5   9   3   8   0  10  24   1  16  21 184   8
       20  10   0   1   0   0   0   1   1   0   1   0   1   0   1   1   1   4   0   1  90
>
```

まとめ: Bayes 推論とNB

$$P(h \mid D) = \frac{P(D \mid h)\,P(h)}{P(D)}$$

- 学習アルゴリズムの俯瞰像:
  - ML: P(D|h) の最大化
  - MAP: P(h|D) ∝ P(D|h) P(h) の最大化
  - Posterior mean:
  - Bayes 最適分類器: P(c|D) = ∫ P(c|h)P(h|D) dh
    - 仮説は分布する！
- Gaussian ノイズ下の回帰:
  ⇔ 二乗誤差の最小化
- 二値事象の確率の学習
  ⇔ cross-entropy の最小化
- Naive Bayes: 乱暴な仮説だが実用的
  - 例えば、文書分類