Bayesian Inference and naïve Bayes

Akito Sakurai

Contents

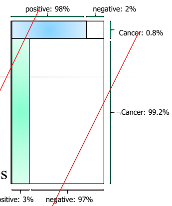
- Bayes theorem
- MAP and ML
- Bayes optimal classifier and Gibbs algorithm
- Prediction of class or probability?
- Naïve Bayes

Bayes Theorem

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

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

Ex. (Mitchell Chap. 6.2)



Suppose we now observe a new patient for whom the lab test returns a positive result. Should we diagnose the patient as having cancer or not?
The test returns a correct positive result in only 98% of the cases in which the disease is actually present and a correct negative result in only 97% of the cases in which the disease is not present. Over the entire population of people only .008 have this disease.

$P(\text{cancer}) = .008$ $\quad P(\neg\text{cancer}) = .992$

$P(+ \mid \text{cancer}) = .98$ $\quad P(- \mid \text{cancer}) = .02$

$P(+ \mid \neg\text{cancer}) = .03$ $\quad P(- \mid \neg\text{cancer}) = .97$

$P(+) = P(+ \mid \text{c'r})\, P(\text{c'r}) + P(+ \mid \neg\text{c'r})\, P(\neg\text{c'r}) = .0376$

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

Ex. (Mitchell Exercise 6.1)

Suppose the doctor decides to order a second laboratory test for the same patient, and suppose the second test returns a positive result as well. What are the posterior probabilities of cancer and ¬cancer following these two tests? Assume that the two tests are independent.

$P(\text{cancer}) = .008$ $\quad P(\neg\text{cancer}) = .992$

$P(+ \mid \text{cancer}) = .98$ $\quad P(- \mid \text{cancer}) = .02$

$P(+ \mid \neg\text{cancer}) = .03$ $\quad P(- \mid \neg\text{cancer}) = .97$

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

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

Basic Probability Formula

Multiplication rule (conditional probability):
(Product rule)

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

Addition rule:
(Sum rule)

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

Law of total probability:

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

Hypothesis selection

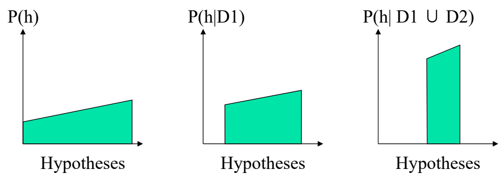
$$P(h\mid D)=\frac{P(D\mid h)\,P(h)}{P(D)}$$

P(h) = prior probability of a hypotheses *h*
P(D) = prior probability that data *D* will be observed
P(h|D) = probability that *h* holds given that *D* is observed
P(D|h) = probability of observing data *D* given *h*

We can estimate the probability that *h* holds under the condition that the training data *D* is observed.
We can, then, estimate the probability that *D* is sampled under *h*.

Note: The conditional probability does not necessarily reflect a causal relationship, if any.
Note: Is it possible to think of the "probability that a hypothesis holds"

Posterior probabilities



Contents

- Bayes theorem
- MAP and ML
- Bayes optimal classifier and Gibbs algorithm
- Prediction of class or probability?
- Naïve Bayes

MAP

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

Finding the most probable hypothesis $h \in H$ given the observed training data D should be most interesting.

<u>*Maximum a posteriori hypothesis*</u> 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

Suppose that $P(h_i) = P(h_j)$ for any i,j, we get Maximum Likelihood (ML) hypothesis

$$h_{ML} = \underset{h\in H}{\operatorname{argmax}} P(D|h)$$

Compare it with:

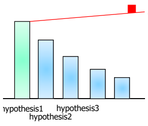
$$h_{MAP} = \underset{h\in H}{\operatorname{argmax}} P(D|h)P(h)$$

An interpretation of ML

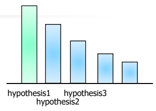
- In the real world, the prior distribution is thought to be unknown, incomputable, or non-existent.
  - E.g., does a prior distribution exist for words in documents? Doesn't it differ in age groups, social background, and others.
- If the existence of a prior distribution is questionable, likelihood-maximization is a reasonable choice.

Most probable classification

- So far, we have obtained the most probable hypothesis given D (h_{MAP}).
- How about most probable class of a sample?
  - $h_{MAP}(x)$ does not predict the most probable class.
    - What is the most probable classification of x ?
    - 3 hypotheses: $P(h_1|D)=0.4$, $P(h_2|D)=0.3$, $P(h_3|D)=0.3$
    - Predictions for a sample: $h_1(x)=+$, $h_2(x)=-$, $h_3(x)=-$



Bayes optimal classification



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

Note: Bayes optimal classifier must not be in H.

Note: Many papers report that it works well; but when we tried we often found no improvement in accuracy compared to MAP or ML. Why does this happen?

Note: Feasible? Doesn't it take time for computation?

Ex. (Mitchell Chap. 6.7)

$$\begin{array}{lll} P(h_1 \mid D) = .4 & P(-\mid h_1) = 0 & P(+\mid h_1) = 1 \\ P(h_2 \mid D) = .3 & P(-\mid h_2) = 1 & P(+\mid h_2) = 0 \\ P(h_3 \mid D) = .3 & P(-\mid h_3) = 1 & P(+\mid h_3) = 0 \end{array}$$

Therefore:
$$\sum_{h_i \in H} P(+|h_i)P(h_i | D) = .4$$
$$\sum_{h_i \in H} P(-|h_i)P(h_i | D) = .6$$

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

Contents

- Bayes theorem
- MAP and ML
- Bayes optimal classifier and Gibbs algorithm
- Prediction of class or probability?
- Naïve Bayes

Bayes optimal classifier

- Suppose that D={x_1,…,x_n } is observed from a distribution P(X;θ) with the parameter θ. We want to estimate y for an unseen x given D.
- Method 1: Estimate θ and then predict by P(X;θ)
  - MLE (max. likelihood) $\theta_{MLE} = \arg\max P(D|\theta)$
  - MAP (max. a posteriori) $\theta_{MAP} = \arg\max P(D|\theta)P(\theta)$
  - Expectation (posterior mean)
    $$\hat{\theta} = \int \theta P(\theta|D)d\theta = \int \theta P(D|\theta)P(\theta)/P(D)d\theta$$
- Method 2: without estimating the parameter θ.
  $$P(Y,\theta|D) = P(Y,D|\theta)P(\theta)/P(D)$$
  $$\Longrightarrow \quad P(Y|D) = \int P(Y,D|\theta)P(\theta)/P(D)d\theta$$

Basic ideas of Bayesian inference

- Bayesian view is that we can measure uncertainty, even if there are not a lot of examples
  - What is the probability that a debut team will win the championship league this year?
    - Cannot do this with a frequentist approach
  - What is the probability that a newly minted particular coin will come up as heads?
  - Without much data we utilize our initial belief as the prior
- But as more data comes available we transfer more of our belief to the data (likelihood)
- With all the data, we do not consider the prior at all
- Belief is coded as a probability distribution

An example: basic ideas

- Assume that we want to infer the mean μ of a random variable x where the variance σ^2 is known and we have not yet seen any data
- $P(\mu|D,\sigma^2) = P(D|\mu,\sigma^2)P(\mu)/P(D) \propto P(D|\mu,\sigma^2)P(\mu)$
- A Bayesian would want to represent the prior μ_0 and the likelihood μ as parameterized distributions (e.g. Gaussian, multinomial, uniform, etc.)
- Let's assume a Gaussian just here
- Since the prior is a Gaussian we would like to multiply it by whatever the distribution of the likelihood is in order to get a posterior which is also a parameterized distribution specifically Gaussian

Conjugate Priors

- $P(\mu|D, \sigma_0^2) = P(D|\mu)P(\mu)/P(D) \propto P(D|\mu)P(\mu)$
- If the posterior is the same distribution as the prior after the multiplication, then we say the prior and posterior are *conjugate* distributions and the prior is a conjugate prior for the likelihood
- In the case of a known variance and a Gaussian prior we can use a Gaussian likelihood and the product (posterior) will also be a Gaussian
- If the likelihood is multinomial then we would need to use a Dirchlet prior and the posterior would be a Dirchlet

Discrete Conjugate Distributions

Discrete distributions [edit]

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

From Wikipedia

Continuous Conjugate Distribution (1)

Continuous distributions [edit]

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters	Interpretation of hyperparameters	Posterior predictive
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	$\mathcal{N}(\tilde{x}\mid\mu_0',\sigma_0^{2\prime}+\sigma^2)$
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	$\mathcal{N}\left(\tilde{x}\mid\mu_0',\frac{1}{\tau_0'}+\frac{1}{\tau}\right)$
Normal with known mean μ	σ^2 (variance)	Inverse gamma	α, β	$\alpha+\frac{n}{2},\ \beta+\frac{\sum_{i=1}^n(x_i-\mu)^2}{2}$	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)	$t_{2\alpha'}(\tilde{x}\mid\mu,\sigma^2=\beta'/\alpha')$
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	$t_{\nu'}(\tilde{x}\mid\mu,\sigma_0^{2\prime})$
Normal with known mean μ	τ (precision)	Gamma	α, β	$\alpha+\frac{n}{2},\ \beta+\frac{\sum_{i=1}^n(x_i-\mu)^2}{2}$	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)	$t_{2\alpha'}(\tilde{x}\mid\mu,\sigma^2=\beta'/\alpha')$
Normal	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	$t_{2\alpha'}\left(\tilde{x}\mid\mu',\frac{\beta'(\nu'+1)}{\nu'\alpha'}\right)$
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	$t_{2\alpha'}\left(\tilde{x}\mid\mu',\frac{\beta'(\nu'+1)}{\alpha'\nu'}\right)$

Wikipediaより

Continuous Conjugate Distribution (2)

Multivariate normal with known covariance matrix Σ	μ (mean vector)	Multivariate normal	$\boldsymbol\mu_0, \boldsymbol\Sigma_0$	$\left(\boldsymbol\Sigma_0^{-1}+n\boldsymbol\Sigma^{-1}\right)^{-1}\left(\boldsymbol\Sigma_0^{-1}\boldsymbol\mu_0+n\boldsymbol\Sigma^{-1}\bar{\mathbf{x}}\right),\ \left(\boldsymbol\Sigma_0^{-1}+n\boldsymbol\Sigma^{-1}\right)^{-1}$; $\bar{\mathbf{x}}$ 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 $\boldsymbol\mu_0$	$\mathcal{N}(\tilde{\mathbf{x}}\mid\boldsymbol\mu_0',\boldsymbol\Sigma_0'+\boldsymbol\Sigma)$
Multivariate normal with known precision matrix Λ	μ (mean vector)	Multivariate normal	$\boldsymbol\mu_0, \boldsymbol\Lambda_0$	$\left(\boldsymbol\Lambda_0+n\boldsymbol\Lambda\right)^{-1}\left(\boldsymbol\Lambda_0\boldsymbol\mu_0+n\boldsymbol\Lambda\bar{\mathbf{x}}\right),\ \left(\boldsymbol\Lambda_0+n\boldsymbol\Lambda\right)$; $\bar{\mathbf{x}}$ is the sample mean	mean was estimated from observations with total precision (sum of all individual precisions) $\boldsymbol\Lambda$ and with sample mean $\boldsymbol\mu_0$	$\mathcal{N}\left(\tilde{\mathbf{x}}\mid\boldsymbol\mu_0',(\boldsymbol\Lambda_0'^{-1}+\boldsymbol\Lambda^{-1})^{-1}\right)$
Multivariate normal with known mean μ	Σ (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 $\boldsymbol\Psi$	$t_{\nu'-p+1}\left(\tilde{\mathbf{x}}\mid\boldsymbol\mu,\frac{1}{\nu'-p+1}\boldsymbol\Psi'\right)$
Multivariate normal with known mean μ	Λ (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 $\mathbf{V}^{-1}$	$t_{\nu'-p+1}\left(\tilde{\mathbf{x}}\mid\boldsymbol\mu,\frac{1}{\nu'-p+1}\mathbf{V}'^{-1}\right)$
Multivariate normal	μ (mean vector) and Σ (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 κ_0 observations with sample mean $\boldsymbol\mu_0$; covariance matrix was estimated from ν_0 observations with sample mean $\boldsymbol\mu_0$ and with sum of pairwise deviation products $\boldsymbol\Psi=\nu_0\boldsymbol\Sigma_0$	$t_{\nu_0'-p+1}\left(\tilde{\mathbf{x}}\mid\boldsymbol\mu_0',\frac{\kappa_0'+1}{\kappa_0'(\nu_0'-p+1)}\boldsymbol\Psi'\right)$
Multivariate normal	μ (mean vector) and Λ (precision matrix)	normal-Wishart	$\boldsymbol\mu_0, \kappa_0, \nu_0, \mathbf{V}$	$\frac{\kappa_0\boldsymbol\mu_0+n\bar{\mathbf{x}}}{\kappa_0+n},\ \kappa_0+n,\ \nu_0+n,\ \left(\mathbf{V}^{-1}+\mathbf{C}+\frac{\kappa_0 n}{\kappa_0+n}(\bar{\mathbf{x}}-\boldsymbol\mu_0)(\bar{\mathbf{x}}-\boldsymbol\mu_0)^T\right)^{-1}$; $\bar{\mathbf{x}}$ 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 κ_0 observations with sample mean $\boldsymbol\mu_0$; covariance matrix was estimated from ν_0 observations with sample mean $\boldsymbol\mu_0$ and with sum of pairwise deviation products $\mathbf{V}^{-1}$	$t_{\nu_0'-p+1}\left(\tilde{\mathbf{x}}\mid\boldsymbol\mu_0',\frac{\kappa_0'+1}{\kappa_0'(\nu_0'-p+1)}\mathbf{V}'^{-1}\right)$

From Wikipedia

Continuous Conjugate Distribution (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	k (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	a, b	$a+n,\ b+\sum_{i=1}^n x_i^\beta$	a observations with sum b of the β'th power of each observation	
Log-normal with known precision τ	μ (mean)	Normal	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	α, β	$\alpha+\frac{n}{2},\ \beta+\frac{\sum_{i=1}^n(\ln x_i-\mu)^2}{2}$	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	$\alpha+n,\ \beta+\sum_{i=1}^n x_i$	α observations that sum to β	$\operatorname{Lomax}(\tilde{x}\mid\beta',\alpha')$ (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	$\operatorname{CG}(\tilde{\mathbf{x}}\mid\alpha,\alpha_0',\beta_0')=\beta'(\tilde{\mathbf{x}}\mid\alpha,\alpha_0',1,\beta_0')$
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)	$\propto\frac{a^{\alpha-1}\beta^{\alpha c}}{\Gamma(\alpha)^b}$	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	α (shape), β (inverse scale)	$\propto\frac{p^{\alpha-1}e^{-\beta q}}{\Gamma(\alpha)^r\beta^{-\alpha s}}$	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	

From Wikipedia

An illustrative example of Bayes inference

- Prior dist.: $P(\mu) = N(\mu \mid \mu_0, \sigma_0^2)$
- Posterior dis.: $P(\mu \mid D) = N(\mu \mid \mu_N, \sigma_N^2)$

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

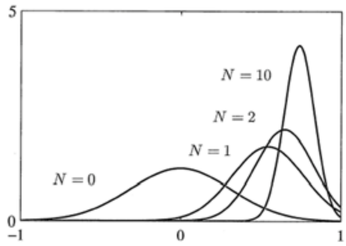
$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

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

- What we believe moves from the prior distribution to data

An illustrative example of Bayes inference

An illustration of Bayesian Inference for the location parameter μ of a Gaussian distribution assuming the variance is given. The curves show the prior distribution of μ (the curve labelled $N = 0$) (this, too, is a Gaussian), along with the posterior distribution when increasing N. The data points are generated from a Gaussian of location and variance parameters 0.8 and 0.1 respectively, and the prior is set to have location parameter 0. Also in the prior distribution and the likelihood function, the true variance is known.



An example of Bayes inference

- In this example, if the mean is known and the variance is unknown, the conjugate prior is the inverse-Gamma.
  - If precision (inverse of variance) is used, the conjugate prior is gamma distribution.
- If mean and variance is unknown, the conjugate prior is normal-inverse-gamma (a combination of normal and inverse-gamma distribution).
- A generalization of this for multivariate case is distribution for multiple dimensions is the normal-inverse-Wishart distribution.

An example of Bayes inference

- $P(\mu,\sigma^2 \mid D) = P(D \mid \mu,\sigma^2)P(\mu,\sigma^2)/P(D)$
  $\propto P(D \mid \mu,\sigma^2)\, P(\mu \mid \sigma^2)P(\sigma^2)$
- prior: $P(\mu \mid \sigma^2) = N(\mu \mid \mu_0, \sigma^2/k_0)$,
  $P(\sigma^2) = \mathrm{IG}(\sigma^2 \mid r_0/2, s_0/2)$
  $$N\_IG(\mu, \sigma^2 | \mu_0, k_0, r_0, s_0)$$
- posterior: $P(\mu \mid \sigma^2, D) = N(\mu \mid \mu_N, \sigma^2/k_N)$,
  $P(\sigma^2) = \mathrm{IG}(\sigma^2 \mid r_N/2, s_N/2)$
  $$N\_IG(\mu, \sigma^2 | \mu_N, k_N, r_N, s_N)$$

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

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

Gibbs classifier

1. Select a hypothesis randomly according to P(h|D)
2. Classify a new example following the h

Good news: If a hypothesis is randomly sampled from P(h) ,

$E[\mathrm{error}_{\mathrm{Gibbs}}] \leq 2E[\mathrm{error}_{\mathrm{BayesOptimal}}]$

(See "Mitchell Machine Learning Chap. 6.8")

Effective when there are so many hypothesis that a Bayes optimal is hard to calculate and we repeat the inferences

Contents

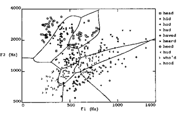
- Bayes theorem
- MAP and ML
- Bayes optimal classifier and Gibbs algorithm
- Prediction of class or probability?
- Naïve Bayes

Two types of target values

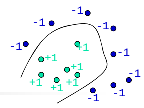
- Categorical val.: classification problem
  - To divide (explanatory variable) space
    - Boundaries are to be obtained
  - Category ↔ value range
- Continuous val.: regression
  - Discrete values, too

When continuous func. is used, these are the same.
(1) e.g. 0 value-set, 1 value-set etc. are boundaries
(2) e.g. non-integer value area represents a category

Note: discrete function is not easily approximated by continuous one, so that regression framework is not straightforwardly applied to it



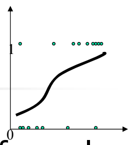
Note



- For categorical values (assuming searching for continuous functions that assumes 0 on boundaries)
  - Close to 0 = close to boundaries = not sure
  - Suppose that confidence level is represented by a real value between 0 and 1, it is, in the sub-area of a category,
    - in the middle = confident = close to 1,
    - close to boundary = not sure = close to 0,

    then the framework is of regression
    - The values are ids of category

Another viewpoint



- Probability and the number of samples
  - If we use probability, not definite value, to represent the level of belongingness to a class, we could suppose that the frequency of samples reflects the probability
- Confidence level and the # of samples
  - Confidence level of being in a subarea for a point is considered to be proportional to the number of samples around the point.

In summary

<table>
<tr><th>Target value</th><th colspan="2">Categorical value</th><th>Continuous value</th></tr>
<tr><td rowspan="4">Concept and method</td><td colspan="2">Find boundary to minimize the # of errors</td><td rowspan="4">Regression of target values (minimize average errors)</td></tr>
<tr><td rowspan="2">Regression of Categorical id values</td><td>Rounded outputs are considered to be id</td></tr>
<tr><td>Fractional part of outputs is confidence</td></tr>
<tr><td colspan="2">Estimate distributions considering inputs as samples from a population (density estimation)</td></tr>
</table>

How to understand NN outputs

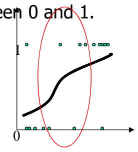
Category id for classification

Although classification class is taught, output values is considered to be probability.

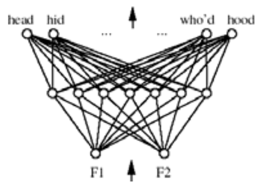
Value for regression

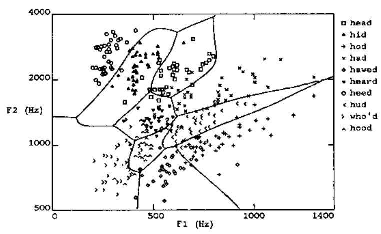
Although classification class is taught, output value is rounded to be considered as category id.

When a standard sigmoid function or the softmax is used as the final activation function, outputs are between 0 and 1.



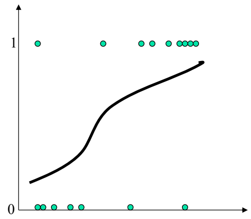
Class areas and boundaries





Learn to predict probability

Why do NNs learn probability, although just classes are taught?

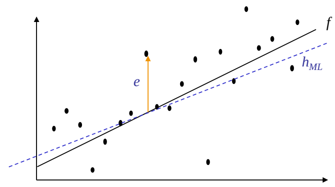


In online learning, frequency is learning by counting.

Is it similar to logistic regression?

Regression to learn reals

By the way, what is regression?



Statistical interpretation

Training samples: <x_i,d_i> where

$d_i = f(x_i) + e_i$

e_i is noise = prob. var. distributed to a normal distribution (iid) of mean=0 and finite deviation

iid=independent, identically distributed

Then:

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

Statistical interpretation

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

Squared error is not appropriate to predict probability

- Ex. Learning survival rate from data

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

d_i is 0 or 1 (or probability belonging to a class)

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

Contents

- Bayes theorem
- MAP and ML
- Bayes optimal classifier and Gibbs algorithm
- Prediction of class or probability?
- Naïve Bayes

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

知的情報処理の復習

Naïve Bayes classifier

- Since (although?) simple, it is wellknown
  - More accurate than expected, although simple
  - Fast as is expected, since simple
- Bayes Theorem + Assumption *conditional independence*
  - The assumption hardly holds in the real world
  - In the real world, though, it works well
- Successful applications:
  - Text classification,
  - Diagnosis, and many others

Naïve Bayes is not a Bayesian

Difficulty in Bayes classifier

- Recall that for a set of attributes $<a_1,\ldots,a_n>$ of x, to infer the class that x belongs

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

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

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

- Difficulty: Huge dataset is required to infer $P(a_1 \ldots a_n | c_j)$, since there are huge number of parameters ($\Pi|A_i|$) (for two value attributes, 2^n parameters for n attributes)

Naïve Bayes classifier

- Naïve Bayes assumption: attributes are mutually independent when the class is given
  - $P(a_1,\ldots,a_n|c_j) = P(a_1|c_j)\, P(a_2|c_j) \ldots P(a_n|c_j)$
  - *conditional independence* (given the class)
  - Reduces the number of parameters to infer: $\Pi|A_i|$ (=O(2^n)) $\rightarrow$ $\Sigma|A_i|$ (=O(n))
- Under this assumption, c_{MAP} becomes

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

Naïve Bayes: an Algorithm

Training (for a set of instances)

Estimate the probability that an instance x belongs to a class c_j

$P^\wedge(c_j) = P(c_j)$'s estimator

Estimate the probability that the i –th attribute value of an instance x belonging to the class c_j is a_i .

$P^\wedge(a_i|c_j)$ = $P(a_i|c_j)$'s estimator

Class(x)

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

Naïve Bayes: Estimation

- How can we estimate $P(c_j)$ and $P(a_i|c_j)$?
  - A standard method that statistics tells us
    - Use frequency of the samples
    - $P(c)$ is estimated by count(c) / N
    - $P(A|B)$ is estimated by count($A \wedge B$) / count(B)
  - Example: 100 samples. 70 + and 30 –
    - P(+)=0.7 and P(-)=0.3
    - Among 70 positives, in 35 a_1=SUNNY
    - P(a_1=SUNNY|+)=0.5

Example: Play Tennis

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	No	No
Sunny	Hot	High	Yes	No
Overcast	Hot	High	No	Yes
Rainy	Mild	High	No	Yes
Rainy	Cool	Normal	No	Yes
Rainy	Cool	Normal	Yes	No
Overcast	Cool	Normal	Yes	Yes
Sunny	Mild	High	No	No
Sunny	Cool	Normal	No	Yes
Rainy	Mild	Normal	No	Yes
Sunny	Mild	Normal	Yes	Yes
Overcast	Mild	High	Yes	Yes
Overcast	Hot	Normal	No	Yes
Rainy	Mild	High	Yes	No

There are two classes: to play tennis (Play=Yes) and not to play tennis (Play=No)

Please infer if on the following day they played tennis or not

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

From Tom Mitchell's Machine Learning

A solution

- For the *PlayTennis*, and a new instance
  <Outlook=sunny, Temp=cool, Humid=high, Windy=true>
- We want to calculate:

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

- $\hat{P}(Y)\hat{P}(sunny \mid Y)\hat{P}(cool \mid Y)\hat{P}(high \mid Y)\hat{P}(true \mid Y) = 0.0053$
  $\hat{P}(N)\hat{P}(sunny \mid N)\hat{P}(cool \mid N)\hat{P}(high \mid N)\hat{P}(true \mid N) = 0.0206$

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

Naïve Bayes: Conditional independence

- Is it necessary?
- What happens if the assumption does not hold?
  - i.e. if $P(a_1,\ldots,a_n|c_j) \neq P(a_1|c_j)\, P(a_2|c_j)\ldots P(a_n|c_j)$
- If the following (weak) condition holds, the prediction is the same as Bayes classifier:

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

- But, the probability obtained in the prediction happens to be unrealistically close to 0 or 1

Naïve Bayes: a Problem

- What happens if an attribute value a_i is not observed for a class c_j ?
  - Estimator of $P(a_i|c_j)=0$ because $\text{count}(a_i \wedge c_j) = 0$
  - Big impacts: if this is 0, any products are 0 !
- A solution: use Laplace correction.
  - $\hat{P}(a_i \mid c_j) = \dfrac{n_c + mp}{n+m}$
  - n : # of training samples for $c = c_j$
  - n_c : # of training samples for $c = c_j$ and $a = a_j$
  - p : prior probability (estimator) $P^\wedge(a_i|c_j)$ (uniform distribution is common)
  - m : pseudo-count (commonly the number of attribute values)

  m=1 is another choice which works better in many cases, too

Note: Laplace correction

- (in parameter estimations from frequency) supposing a prior distribution for the parameter, obtain a MAP estimator.
- Beta distribution is the prior:
  $f(x;\alpha,\beta)=x^{\alpha-1}(1-x)^{\beta-1}/B(\alpha,\beta)$
- The posterior mean of the parameter is the Laplace correction. If the likelihood is a result of a Bernoulli trial: $\hat{\theta} = (n_0+\alpha)/((n_0+n_1)+\alpha+\beta)$

Note: smoothing

- In an estimation of statistical model, assigning a small probability to events that did not occur is called smoothing
  http://www.jaist.ac.jp/project/NLP_Portal/doc/glossary/index.html
- In natural language processing, frequencies of a word or a sequence of n words (n-gram) are often used. When n grows, n-gram becomes scarce, i.e., many n-grams do not occur. To solve the problem many techniques were invented.
  - Laplace smoothing (additive smoothing)
  - Linear interpolation
  - Good-Turing smoothing
  - Katz smoothing
  - Church-Gale smoothing
  - Witten-Bell smoothing
  - Kneser-Ney smoothing
  - ....
  - Hierarchical Pitman-Yor language model

Text classification

- Texts classification:
  - Classifying documents (mail, news, web pages, etc. or a paragraph, a sentence, etc.)
  - Classifying e-mails into spam or not.
  - Classifying blogs into splog or not
  - Classifying news into interesting or not (to a person)
  - Classifying reviews of a product into groups of good reputation or not
  - Classifying reviews into trustable or not
  - Classifying open ended questions for questionnaire surveys
  - Classifying Q and A's at a call-center.
- Naïve Bayes works well
  - How to apply Naïve Bayes ?
  - Point: How to represent a sample (i.e. document), attributes?

Document representation

Bag-of-words
- Document as a vector of frequency of words in it
  - "Bag" implies discarding positions where the word occurs, and
  - disregarding the sequences (contexts) of a word
  - i.e. if keio, gijuku, and university are words, there would be no difference between keio gijuku university, keio university giju, and gijuku keio university
- "what are words" is important, which should not differ among documents.
  - In English, "dog" and "dogs" should be treated as the same
- Ignore words not relevant to classification
  - In Japanese, particles such as ha, ga, mo, ya, etc are the ones
  - In English, prepositions
  - The words that have syntactic function but have no meaning are called functional words.
- Ignore words that are close to noise
  - Very low frequent words such as appearing just once.

Document representation (cntd.)

- Representation itself is like naïve Bayes
  - Because representation is not inference, it is not naïve Bayes, but it really looks like naïve Bayes.
  - Probability of the occurrence of a document is formulated in naïve Bayes fashion.
  - Suppose that for each class of documents, the probability that a specific word occurs in a document is known as $P(w_1|c_j)$, $P(w_2|c_j)$, …, $P(w_n|c_j)$. If $w_1, w_2, \ldots, w_n$ are the words that occur in a document, then the probability that the document occurs is

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

    where TF(w) is the term frequency of a word w in a document doc

出現確率をこう書けば naïve Bayes といえよう

Document classification by Naïve Bayes

- For a document doc,

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

where $TF(w_k,\mathrm{doc})$ = frequency of w_k in doc and Voc is a set of all the words that we consider

- To represent word frequencies in a document, we need Laplace correction. The following estimator is used; where n_j=the number of words in a class c_j , $n_{k,j}$=the number of occurrences of word w_k in class c_j .

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

Twenty News Groups (Joachims 1996)

- 1000 training documents in each group
- Assign new documents to one of newsgroups

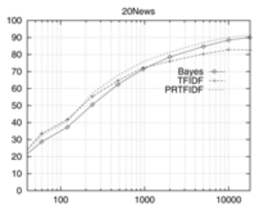
comp.graphics	misc.forsale
comp.os.ms-windows.misc	rec.autos
comp.sys.ibm.pc.hardware	rec.motorcycles
comp.sys.mac.hardware	rec.sport.baseball
comp.windows.x & rec.sport.hockey	rec.sport.hockey
alt.atheism	sci.space
soc.religion.christian	sci.crypt
talk.religion.misc	sci.electronics
talk.politics.mideast	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% accuracy of classification
  - Highly frequent 100 word (the and of …) are deleted
    - The words such as functional words, words relatively useless for classification are categorized as stop words and are deleted
  - The words occurring less than 3 times are deleted
  - The words remained: 38,500

Note: the accuracy is overly high.. In every text in 20 Newsgroups has a "subject" field which is very helpful for classification. Although the subject field is now deleted, in the previous works the field might be utilized.



Accuracy vs. # of training data (1/3 is reserved for test)

20 Newsgroups in R

- Naïve Bayes package in R is not appropriate to large dataset.
  - Because the data matrix with naïve implementation becomes huge (in the previous R program, the matrices xy, xy, tt) (2000 rows for documents and 40,000 columns for words).
  - But non-zero entries are small in number, sparse matrix representation is useful.
  - You have to pay for computational overhead.
  - Then let us write an efficient program by ourselves!?

20 Newsgroups: データ

- In the "20 Newsgroups" site:
  - http://people.csail.mit.edu/jrennie/20Newsgroups/
  - Redirected to http://qwone.com/~jason/20Newsgroups/
- There is a preprocessed version:
  - 20news-bydate-matlab.tgz
- We use train.data, train.label, test.data, and test.label
- A program is uploaded to the class web page.
- Only the confusion matrix is in the next slide.
- Accuracy is around 78.2%.

```
> 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   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
>
```

Bayes inference and NB

$$P(h|D)=\frac{P(D|h)P(h)}{P(D)}$$

- Overview of the learning algorithm:
  - ML: maximize P(D|h)
  - MAP: maximize P(h|D) ∝ P(D|h) P(h)
  - Posterior mean:
  - Bayes optimal classifier: P(c|D) = ∫ P(c|h)P(h|D) dh
    - Hypotheses distribute!
- Regression under Gaussian noise:
  ⇔ minimization of mean squared error
- Learning of probability of binary events
  ⇔ minimization of cross-entropy
- Naive Bayes: rough assumption but practical
  - Ex. Document classification