CS 395T Lecture 12 Classification

April 12, 2000

Recap: Probabilistic Learning Formulation

$$P(w_i|x) \alpha P(x|w_i)P(w_i), i = 1, ..., c \quad (1)$$

Boils down to modeling $P(x|wi)$

$->$ Naive Bayes / Bayesian Networks (Discrete *)
$->$ Maximum Likelihood Classifiers (Real-values *)

Naive Bayes Example

On Yahoo's site, we looked at "Yahoo! Education" and "Yahoo! Sports" classes and searched for the words "grade" and "score". Following were the results:

	Word	
Class	grade	score
Yahoo! Education	2423	113
Yahoo! Sports	143	848
Others	587	887
Total	3153	1848

N = Total # documents
Ne = # documents in "Education" class, 113,000
Ns = # documents in "Sports" class, 160,000
No = # documents in "Others" class
The words "grade" and "score" both occurred in only 8 documents.

What is the probability that a document (web page) that contains the word "grade" and "score" belongs to "Yahoo! Education" on "Yahoo! Sports"?

$$P(Education|grade, score) = \frac{P(Education).P(grade, score|Education)}{P(grade, score)} \quad (2)$$

$$= \frac{P(Education)P(grade|Education)P(score|Education)}{P(grade, score)} \quad (3)$$

$$
\begin{aligned}
&= \frac{\frac{N_e}{N}\frac{2423}{N_e}\frac{113}{N_e}}{\frac{8}{N}} \\
&\approx 0.30
\end{aligned}
$$

Equation (2) represents Bayes Rule and equation (3) represents Naive Bayes. Similarly,

$$
\begin{aligned}
P(Sports|grade, score) &= \frac{P(Sports)P(grade|Sports)P(score|Sports)}{P(grade, score)} \\
&= \frac{\frac{N_s}{N}\frac{143}{N_s}\frac{848}{N_s}}{\frac{8}{N}} \\
&\approx 0.09
\end{aligned}
$$

Usual assumption is that $P(x|w_i)$ is a <u>NormalGaussian Distribution</u>

- most naturally occuring distributions are Gaussians
- noise models
- two parameters distribution (for each class)

$$
P(x|\mu, \Sigma) = \frac{1}{(\sqrt{2\pi})^d |\Sigma|^{1/2}} exp[-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)] \tag{4}
$$

μ Mean of Distribution $= E[x]$
Σ Covaciance matrix $= E[(x-\mu)(x-\mu)^T]$, Σ is symmetric
d dimmensionality of data

In practice, x is a $d*1$ vector, μ is $d*1$ and Σ is $d*d$.

$(x-\mu)^T\Sigma^{-1}(x-\mu)$: squared Mahalanobis distance from x to μ, which takes care of dependence and scaling.

<u>Examples</u>:

$$
\Sigma = \sigma^2 I \tag{5}
$$

Special Distribution (all dimensions are indenpendnet and have same variance).

$$
\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \sigma_d^2 \end{bmatrix} \tag{6}
$$

Hyperellipsoids with semi-axes parallel to the candidate axes.

In general, Σ may be a full, positive-define matrix.

For example, in the 2*2 case, Σ has 2 eigenvalues & eigenvectors:

$$\begin{aligned} \Sigma\mu_1 = & \quad \lambda_1\mu_1 \\ \Sigma\mu_2 = & \quad \lambda_2\mu_2 \end{aligned}$$

No. of parameters per class (remember there is 1 Gaussian per class)

$$\begin{array}{lll} \sigma^2 I & -> & 1 \text{ (Spherical)} \\ \text{diagonal} & -> & d \text{ (axis parallel hyperellipsoid)} \\ \text{full} & -> & d + \frac{d(d-1)}{2} = \frac{d(d+1)}{2} \text{ (Symmetric matrix)} \end{array}$$

A. Using Known Gaussian Distributions for Classification

Suppose we know the c class distributions:

$$\begin{aligned} P(x|w_i) &= P(x|\mu_i, \Sigma_i) \\ &= \frac{1}{(\sqrt{2\pi})^d |\Sigma_i|^{\frac{1}{2}}} exp[-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)] \end{aligned}$$

Then the decision(MAP) rule reduces to: Given a new point x classify it such that

$$\begin{aligned} w(x) &= arg\ max_{i=1,\ldots,c} P(wi|x) \qquad (7) \\ &= arg\ max_{1\leq i\leq c} \frac{P(x|w_i)P(w_i)}{P(x)} \quad Byyes\ rule \\ &= arg\ max_{1\leq i\leq c} P(x|w_i)P(w_i) \quad Ignore\ the\ common\ denominator \\ w(x) &= arg\ max_{1\leq i\leq c} \ln P(x|w_i) + \ln P(x|w_i) \quad Lodarithm\ is\ a\ monotonic\ function \end{aligned}$$

We call $y_i(x) = \ln P(w_i) + \ln P(x|wi)$ as the Discriminant function. Decision rule: For any new point x, compute $y_i(x)$ for every i and assign class label

$$w(x) = arg\ max_{1\leq i\leq c} y_i(x) \qquad (8)$$

B. Parameter Estimation

Parameters to be estimated? $\{\mu_i, \Sigma_i\}$ Given $X = X_1 \cup X_2 \cup \ldots X_c$, where X_i is set of points in i^{th} class. Estimate $\{\mu_i, \Sigma_i\}$ using MAXIMUM LIKELIHOOD ESTIMATE (MLE).

Maximum Likelihood Classifier

Step1: Given $X = X_1 U X_2 U ... U X_C$ (given C classes, where X_i is the set of points in the ith class.)

Step2: Compute $U_i, (\sum)_i$ using Maximum Likelihood Estimate. This gives $p(x|w_i)$.

Step3. Given $p(x|w_i)$ and a new data point x_{new}, compute the discriminant function

$$y_i(x_{new}) = \ln p(w_i) + \ln p(x_{new}|w_k)$$

Step4. Apply MAP to determine class:

$$w(x_{new}) = arg \max_{1\leq i\leq c} \ y_i(x_{new})$$

Discriminant Function

$$y_i(x) = \ln P(w_i) + \ln[\frac{1}{(\sqrt{2\pi})^d |\Sigma|^{1/2}} \exp[\frac{-1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)]]$$

$$y_i(x) = \ln P(w_i) - \frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(x-\mu_i)^T(\Sigma_i)^{-1}(x-\mu_i) - \frac{d}{2}\ln 2\pi$$

Decision Boundary for 2-class problem

$y_1(x) = y_2(x) \rightarrow$ Locus of all x on the boundary surface

In general, the decision boundary will be Quadratic

Case 1

$\Sigma_k = \Sigma$ (ie. every class has identical covariance matrix)

$$\begin{aligned} y_i(x) &= \ln P(w_i) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i) \\ &= \ln P(w_i) - \frac{1}{2}[x^T\Sigma^{-1}x - \mu_i^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_i + \mu_i^T\Sigma^{-1}\mu_i] \\ &= \ln P(w_i) - \frac{1}{2}\mu_i^T\Sigma^{-1}\mu_i + (\Sigma^{-1}\mu_i)^T x \end{aligned}$$

or

$$y_i(x) = w_i^T x + w_{i,0} \tag{9}$$

where $w_i = \Sigma^{-1}\mu_i, w_{i,0} = -\frac{1}{2}\mu_i^T\Sigma^{-1}\mu_i + \ln P(w_i)$

$\Rightarrow linearDiscriminantFunction$

Linear Decision Boundary

$y_1(x) = y_2(x)$

$\Rightarrow w_1^T x + w_{1,0} = w_2^T x + w_{2,0}$
$\Rightarrow (w_1 - w_2)^T x + (w_{1,0} - w_{2,0}) = 0$
$\Rightarrow (\mu_1 - \mu_2)^T\Sigma^{-1}x + \ln\frac{P(w_1)}{P(w_2)} - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 = 0$
$\Rightarrow w^t x + w_0 = 0, where\ w = (\Sigma)^{-1}(\mu_1 - \mu_2) and$
$w_0 = -\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 + \ln\frac{P(w_1)}{P(w_2)}$

if $\Sigma = \sigma^2 I$ and $P(w_1) = P(w_2)$
$y_1(x) = y_2(x)$

$\Rightarrow (\mu_1 - \mu_2)^T x - \frac{1}{2}\mu_1^T \mu_1 + \frac{1}{2}\mu_2^T \mu_2 = 0$
$\Rightarrow ||x - \mu_1||^2 = ||x - \mu_2||^2$

Hyperplane normal to the line joining μ_1 and μ_2 and passing midway between μ_1 & μ_2.