EE 649 Pattern Recognition

Discrete Classifiers

Ulisses Braga-Neto

ECE Department

Texas A&M University

Main Ideas

- Also known as *multinomial discrimination* or *categorical classification*.
- Predictor Variables X_i can only assume discrete (i.e., finitely many) values.
- The discrete values can be either numeric or nominal (in the case of discrete histogram rule, it does not matter).
- Very important in biology (genomics), psychology, economy, and social sciences.
- Popular in Data Mining.

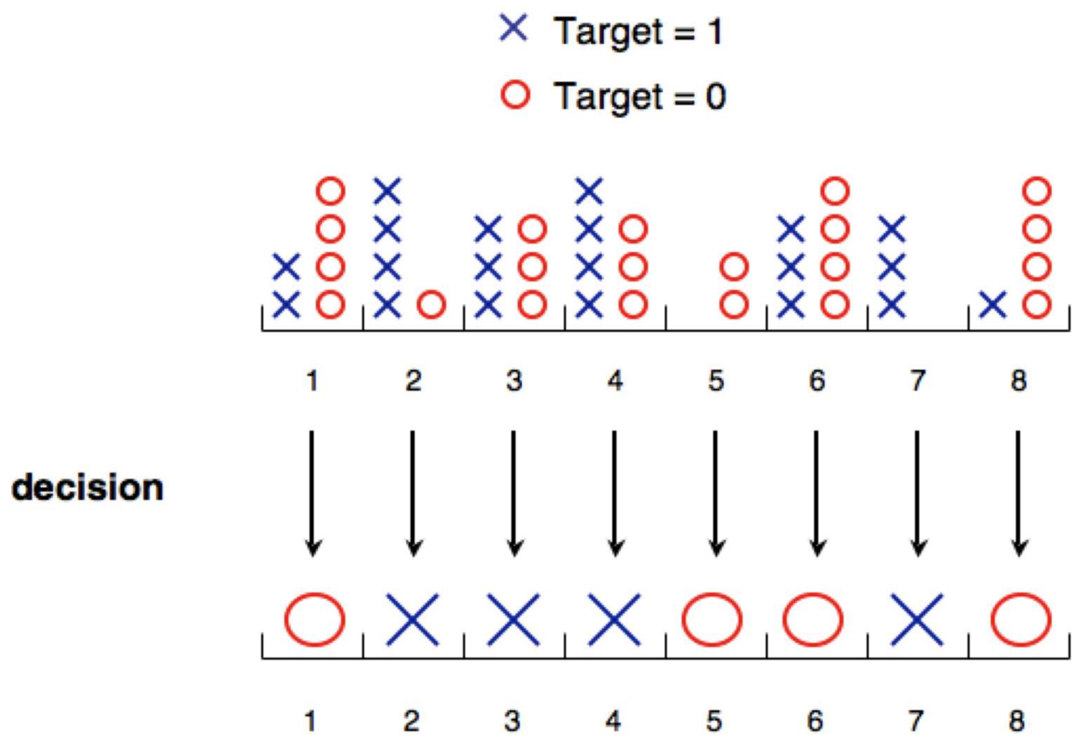
Discrete Histogram Rule

The feature space in discrete classification is a finite grid. A very popular (and natural) tree classifier is thus the discrete histogram rule.

X_1 = color \ X_2 = size	big	medium	small
red		<u>apple = 6</u> grape = 3	
green	<u>watermellons = 8</u>	<u>apple = 3</u> watermellons = 1	<u>grape = 5</u> apple = 1
yellow	<u>grapefruit = 8</u> lemon = 1	lemon = 2 grapefruit = 2	<u>lemon = 5</u>

Discrete Histogram Rule - II

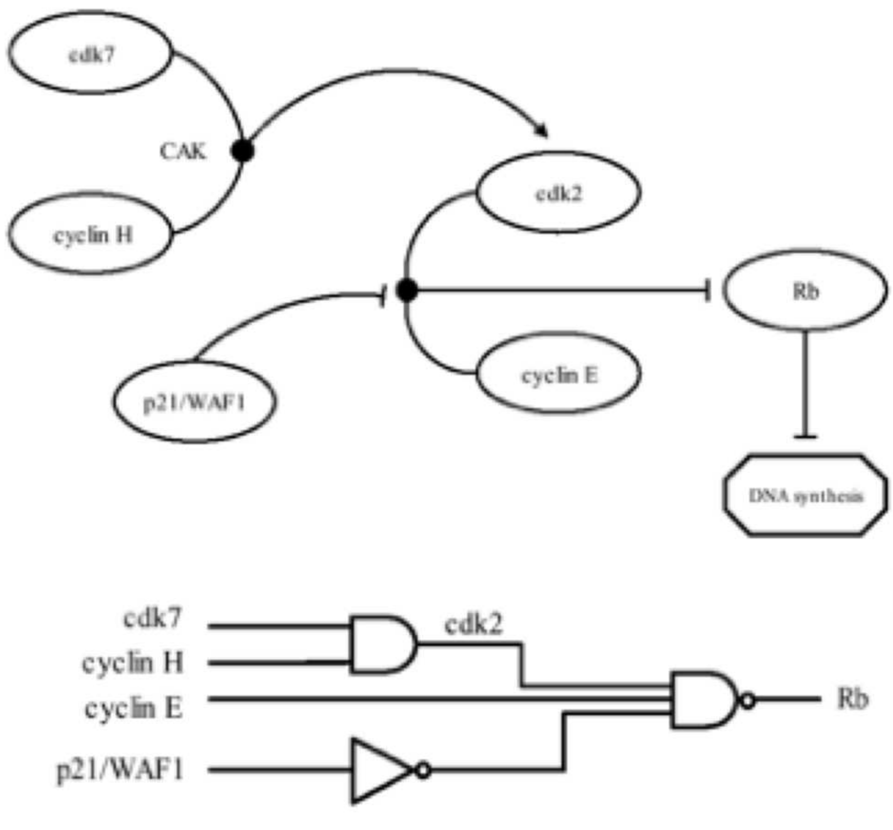
The discrete histogram rule can be seen as “majority voting over bins.”



“Clearly,” the discrete histogram rule is strongly universally consistent (why?)

Example: Gene Regulatory Networks

In this application, the predicting variables and the target are binary gene expressions.



Quantized Gene Expression Data

cdk7	cyc H	cyc E	p21/W	Rb
0	1	0	1	1
0	0	1	0	0
1	1	0	1	1
0	1	0	1	0
1	0	1	0	1
0	1	0	1	0

Mathematical Formulation

There is a feature vector $\mathbf{X} = (X_1, \ldots, X_p)$, consisting of p predictor variables, such that each X_i takes on a finite number b_i of values, and a discrete target variable $Y \in \{0, 1, \ldots, c-1\}$ (we will assume $c = 2$).

The predictors as a group take on values in a finite space of $b = \prod_{i=1}^{p} b_i$ possible "states."

The value b is the number of "bins" into which the data is categorized — it provides a direct measure of the complexity of the classification rule.

Model Parameters

Let $(\mathbf{x}^1, \ldots, \mathbf{x}^b)$ be an arbitrary enumeration of the states. The complete probability structure of the discrete classification problem is specified by $2b+2$ real numbers:

- The class prior probabilities

$$c_0 = P(Y=0)$$
$$c_1 = P(Y=1)$$

- The class-conditional probabilities:

$$p_i = P(\mathbf{X}=\mathbf{x}^i \mid Y=0)\,, \quad i = 1, \ldots, b$$
$$q_i = P(\mathbf{X}=\mathbf{x}^i \mid Y=1)\,, \quad i = 1, \ldots, b$$

These parameters completely determine the joint probability $P(\mathbf{X}=\mathbf{x}^i, Y=j)$ and thus the stochastic problem.

Discrete Bayes Classifier

For $i = 1, \ldots, b$, we have that

$$\begin{aligned} \eta(\mathbf{x}^i) &= P(Y{=}1 \mid \mathbf{X}{=}\mathbf{x}^i) = q_i c_1 / P(\mathbf{X}{=}\mathbf{x}^i) \\ 1 - \eta(\mathbf{x}^i) &= P(Y{=}0 \mid \mathbf{X}{=}\mathbf{x}^i) = p_i c_0 / P(\mathbf{X}{=}\mathbf{x}^i) \end{aligned}$$

Therefore, the Bayes classifier is:

$$\psi^*(\mathbf{x}^i) = \begin{cases} 1, & \eta(\mathbf{x}^i) > 1 - \eta(\mathbf{x}^i) \\ 0, & \text{otw} \end{cases} = \begin{cases} 1, & p_i c_0 < q_i c_1 \\ 0, & \text{otw} \end{cases}$$

with corresponding optimal error rate:

$$\epsilon^* = \sum_{i=1}^{b} \min\{p_i c_0, q_i c_1\}$$

Data-Based Discrete Classification

This is nice, but in practice, we do not know the model parameters c_0, c_1, p_i and q_i, but we only know the sample training data $S_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$.

Let us introduce the observed bin counts:

$$U_i = \#\{\mathbf{X}_j = \mathbf{x}^i \mid Y_j = 0\}, \;\; i = 1, \ldots, b,$$

$$V_i = \#\{\mathbf{X}_j = \mathbf{x}^i \mid Y_j = 1\}, \;\; i = 1, \ldots, b.$$

Let us also define $N_0 = \sum_{i=1}^{b} U_i$ and $N_1 = \sum_{i=1}^{b} V_i$, such that $N_0 + N_1 = n$. Notice that N_0, N_1, U_i and V_i, $i = 1 \ldots, b$ are random variables.

For the purpose of discrete histogram classification, the data S_n can be summarized by U_i and V_i, $i = 1 \ldots, b$ alone.

Discrete Plug-in Rule

Suppose we try to approximate the unknown a-posteriori probability $\eta(i)$ by using the maximum-likelihood sample-based estimates of the model parameters:

$$\widehat{c}_0 = \frac{N_0}{n}, \;\; \widehat{c}_1 = \frac{N_1}{n} \quad \text{and} \quad \widehat{p}_i = \frac{U_i}{N_0}, \;\; \widehat{q}_i = \frac{V_i}{N_1}, \;\; \text{for } i = 1, \ldots, b.$$

Plugging these back in the expression for the Bayes classifier leads to the plug-in classifier:

$$\psi_n(\mathbf{x}^i) \;=\; I_{V_i > U_i} \;=\; \begin{cases} 1, & V_i > U_i \\ 0, & \text{otw} \end{cases}, \qquad i = 1, \ldots, b.$$

which is none other than the discrete histogram classifier! In other words, the discrete histogram rule is the plug-in rule for discrete classification.

Error of Discrete Histogram Classifier

We have that

$$
\begin{aligned}
\epsilon_n &= P(\psi_n(\mathbf{X}) \neq Y) = \sum_{i=1}^{b} P(\mathbf{X}\!=\!\mathbf{x}^i,\, Y\!=\!1\!-\!\psi_n(\mathbf{x}^i)) \\
&= \sum_{i=1}^{b} P(\mathbf{X}\!=\!\mathbf{x}^i \mid Y\!=\!1\!-\!\psi_n(\mathbf{x}^i))P(Y\!=\!1\!-\!\psi_n(\mathbf{x}^i)) \\
&= \sum_{i=1}^{b} \left[p_i c_0 I_{\psi_n(\mathbf{x}^i)=1} + q_i c_1 I_{\psi_n(\mathbf{x}^i)=0}\right] \\
&= \sum_{i=1}^{b} \left[\, c_0 p_i \, I_{V_i > U_i} + c_1 q_i \, I_{U_i \geq V_i} \,\right].
\end{aligned}
$$

Expected Classification Error

Indicator random variables are nice: we have the property $E[I_A] = P(A)$ for any event A. We can take advantage of this to compute the expected error over the sample

$$E[\epsilon_n] = \sum_{i=1}^{b} \left[\, c_0 p_i \, E[I_{V_i > U_i}] + c_1 q_i \, E[I_{U_i \geq V_i}] \,\right]$$

$$= \sum_{i=1}^{b} \left[\, c_0 p_i P(V_i > U_i) + c_1 q_i P(U_i \geq V_i) \,\right]$$

Note: Since $E[\epsilon_n] \to \epsilon^*$ as $n \to \infty$, the discrete histogram rule is consistent (homework).

Expected Classification Error - II

The probability $P(V_i > U_i)$ can be computed by realizing that the pair of random variables (U_i, V_i) is *trinomially* distributed with parameters $(n, c_0 p_i, c_1 q_i)$, i.e.

$$P(U_i = k, V_i = l) = \binom{n}{k, l, n-k-l} (c_0 p_i)^k \, (c_1 q_i)^l \, (1 - c_0 p_i - c_1 q_i)^{n-k-l}$$

for $k, l = 0, \ldots, n$ with $k + l \leq n$. We can then write:

$$P(V_i > U_i) = \sum_{\substack{k,l=0 \\ k<l \\ k+l \leq n}}^{n} P(U_i = k, V_i = l)$$

Example

Zipf model:

$$p_i = \frac{K}{i^\alpha}$$

$$q_i = p_{b-i+1}$$

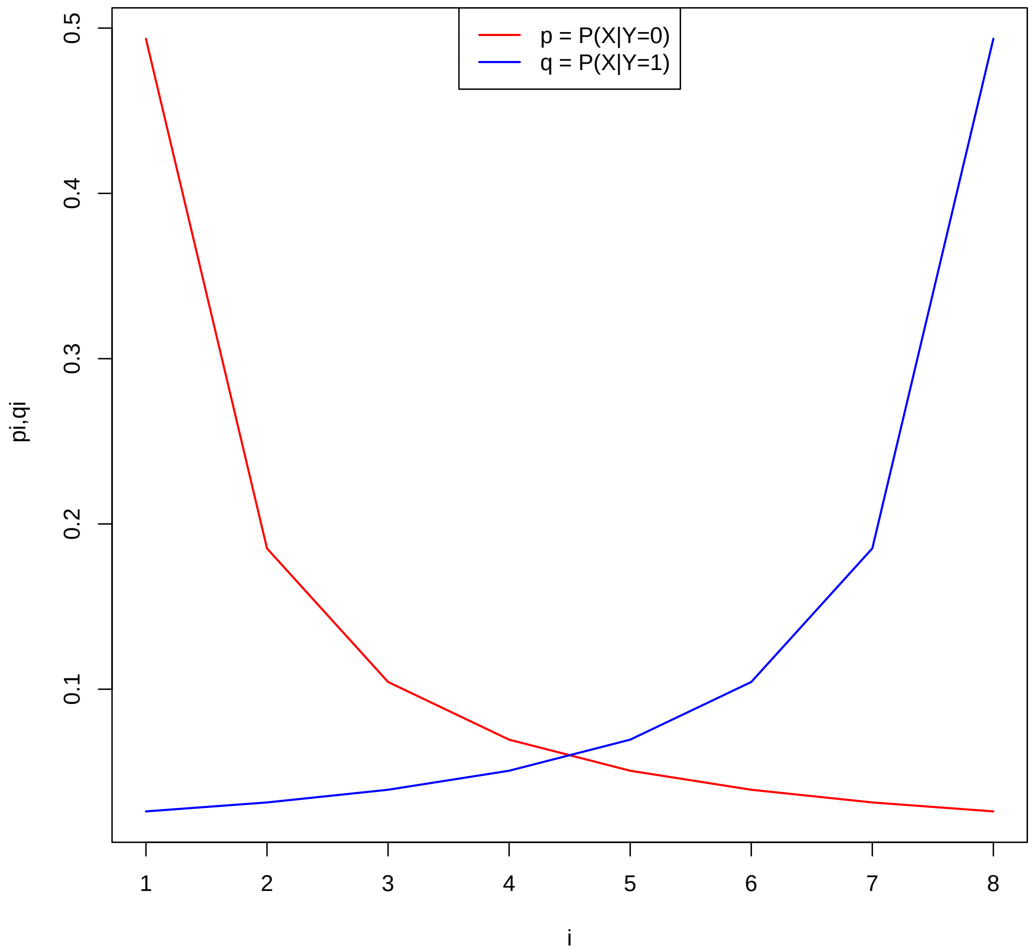
for $i = 1, \dots, b$. Here $\alpha > 0$, and the normalizing constant K is given by:

$$K = \left[\sum_{i=1}^{b} \left(\frac{1}{i^\alpha}\right)\right]^{-1}$$

As $\alpha \to 0$, the distributions become uniform (maximum confusion between classes) whereas, as $\alpha \to \infty$, the distributions become concentrated in single (distinct) bins (maximum discrimination between the classes)

Example - II

Class-conditional distributions for Zipf model with $\alpha = \sqrt{2}$



Example - III

Expected classification error versus sample size for Zipf model with $\alpha = \sqrt{2}$.