Recognition Using Strings

If patterns are represented as strings, it is difficult or impossible to apply metric-based methods.
Why use strings?

- Its natural - gene sequences, textual data
- A convenient description - chain codes for geometric shapes image pixel arrangements, musical notes.

String Recognition Problems

- String matching
- Edit distance between strings - minimal basic operations
- String matching with errors
- String matching with don’t care

String Matching

String matching is a form of template matching. Is *string x* embedded in *string y* or is $x \subseteq y$, where $|x| < |y|$.

$text =$ t g o g o r d x g g o l d r f

$x =$ g o l d

The naive solution is to simply test each substring until the necessary shift is found.

1 15

t g o g o r d x g g o l d r f

g o l d

$\longleftarrow \; shift\ 9 \; \longrightarrow$

$$O(n, m) = O(m(n - m + 1))$$

Boyer-Moore String Matching

1. Match from right-to-left.
2. The Bad-character hueristic suggests a shift that aligns the rightmost occurence of the bad character in *x* with the same character in *text*.
3. The Good-character hueristic suggests a shift that aligns the next occurrence of the good suffix of *x* with the good suffix in *text*.
4. Choose the larger shift.

Boyer-Moore Continued

c	c	b	a	b	a	b	c	c	a	b	c	d	a	b

a	b	c	d	a	b

No match, shift 1.

c	c	b	a	b	a	b	c	c	a	b	c	d	a	b

	a	b	c	d	a	b

ab matches. $F(x) \rightarrow$ shifts 2 to align b with b. $B(x) \rightarrow$ suggests shift 4 to match ab with ab. Take the larger.

Boyer-Moore Continued

c	c	b	a	b	a	b	c	c	a	b	c	d	a	b

					a	b	c	d	a	b				

ab matches. F(x) suggests a shift of 2, B(x) suggests shift 4, and we have a match.

Edit Distance

Given a string, x and a set of classified strings $\mathbf{y}$, the objective is to find the nearest neighbor match for x.
The matching is done by transforming x into $\mathbf{y}_i$ by a series of fundamental operations:

- Substitutions, replace a character in x with a character from $\mathbf{y}$.
- Insertions, insert a character from $\mathbf{y}$ into x.
- Deletions, delete a character from x.

Edit Distance Continued

conjugation $\rightarrow$ confrontation

- Match
- Match
- Match
- Sustitute j for f.
- Sustitute u for r.
- Sustitute g for o.
- Delete n.
- Delete t.
- Match x 5

Total cost is 5, unless the costs are asymmetric.

String Matching with Errors

Perform the matching algorithm, but calculate and save the edit distance for each match.