Feature selection

Devika Subramanian
Comp 470

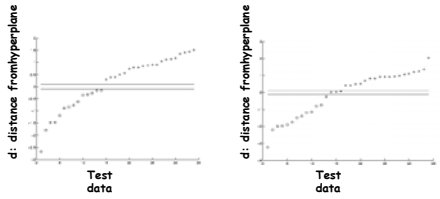
Feature/gene selection

- SVMs as covered in class use all features or genes.
- However, molecular biologists and oncologists believe that only a small number of genes are responsible for particular biological properties.
- When the number of samples is much much smaller than the number of features, over-fitting is very likely. Possible performance improvement (reduction in over-fitting) with fewer features.
- Feature selection is thus a very important problem in classification studies of gene expression data.

SVM with gene selection

AML vs ALL: 40 genes 34/34 correct, 0 rejects.

5 genes 31/34 correct, 3 rejects of which 1 is an error.



Two approaches

- Filter: make an independent assessment based on general characteristics of the data. The feature set is filtered to produce the most promising subset before learning.
- Wrapper: to evaluate the feature subset using the machine learning algorithm that will ultimately be employed for learning. The learning algorithm is wrapped into the feature selection procedure.

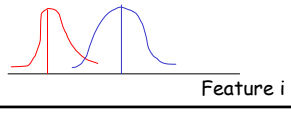
Filter technique: Fisher scores

- For each feature, compute the fisher index

$$Fisher(i) = \left| \frac{\mu_i^+ - \mu_i^-}{(\sigma_i^+)^2 + (\sigma_i^-)^2} \right|$$

μ_i^+ is the mean of the i^{th} feature among the + class

σ_i^+ is the std. dev. of the i^{th} feature among the + class



Wrapper technique: Recursive feature elimination

- Solve the SVM problem and find vector w.
- Rank order elements of w by absolute value.
- Discard features/genes corresponding to the bottom 10% of the values.
- Retrain SVM on reduced gene set and go back to step 2

Example

Dataset	Total Samples	Class 0	Class 1
Leukemia Morphology (train)	38	27 ALL	11 AML
Leukemia Morpholgy (test)	34	20 ALL	14 AML
Leukemia Lineage (ALL)	23	15 B-Cell	8 T-Cell
Lymphoma Outcome (AML)	15	8 Low risk	7 High risk

Dataset	Total Samples	Class 0	Class 1
Lymphoma Morphology	77	19 FSC	58 DLCL
Lymphoma Outcome	58	22 Low risk	36 High risk
Brain Morphology	41	14 Glioma	27 MD
Brain Outcome	50	38 Low risk	12 High risk

Hierarchy of difficulty:

1. Histological differences: normal vs. malignant, skin vs. brain
2. Morphologies: different leukemia types, ALL vs. AML
3. Lineage B-Cell vs. T-Cell, folicular vs. large B-cell lymphoma
4. Outcome: treatment outcome, elapse, or drug sensitivity

Results: part 1

Dataset	Algorithm	Total Samples	Total errors	Class 1 errors	Class 0 errors	Number Genes
Leukemia Morphology (trest) AML vs ALL	SVM	35	0/35	0/21	0/14	40
	WV	35	2/35	1/21	1/14	50
	k-NN	35	3/35	1/21	2/14	10
Leukemia Lineage (ALL) B vs T	SVM	23	0/23	0/15	0/8	10
	WV	23	0/23	0/15	0/8	9
	k-NN	23	0/23	0/15	0/8	10
Lymphoma FS vs DLCL	SVM	77	4/77	2/32	2/35	200
	WV	77	6/77	1/32	5/35	30
	k-NN	77	3/77	1/32	2/35	250
Brain MD vs Glioma	SVM	41	1/41	1/27	0/14	100
	WV	41	1/41	1/27	0/14	3
	k-NN	41	0/41	0/27	0/14	5

Results: part 2

Dataset	Algorithm	Total Samples	Total errors	Class 1 errors	Class 0 errors	Number Genes
Lymphoma LBC treatment outcome	SVM	58	13/58	3/32	10/26	100
	WV	58	15/58	5/32	10/26	12
	k-NN	58	15/58	8/32	7/26	15
Brain MD treatment outcome	SVM	50	7/50	6/12	1/38	50
	WV	50	13/50	6/12	7/38	6
	k-NN	50	10/50	6/12	4/38	5