Issues in Learning

Adrian Barbu

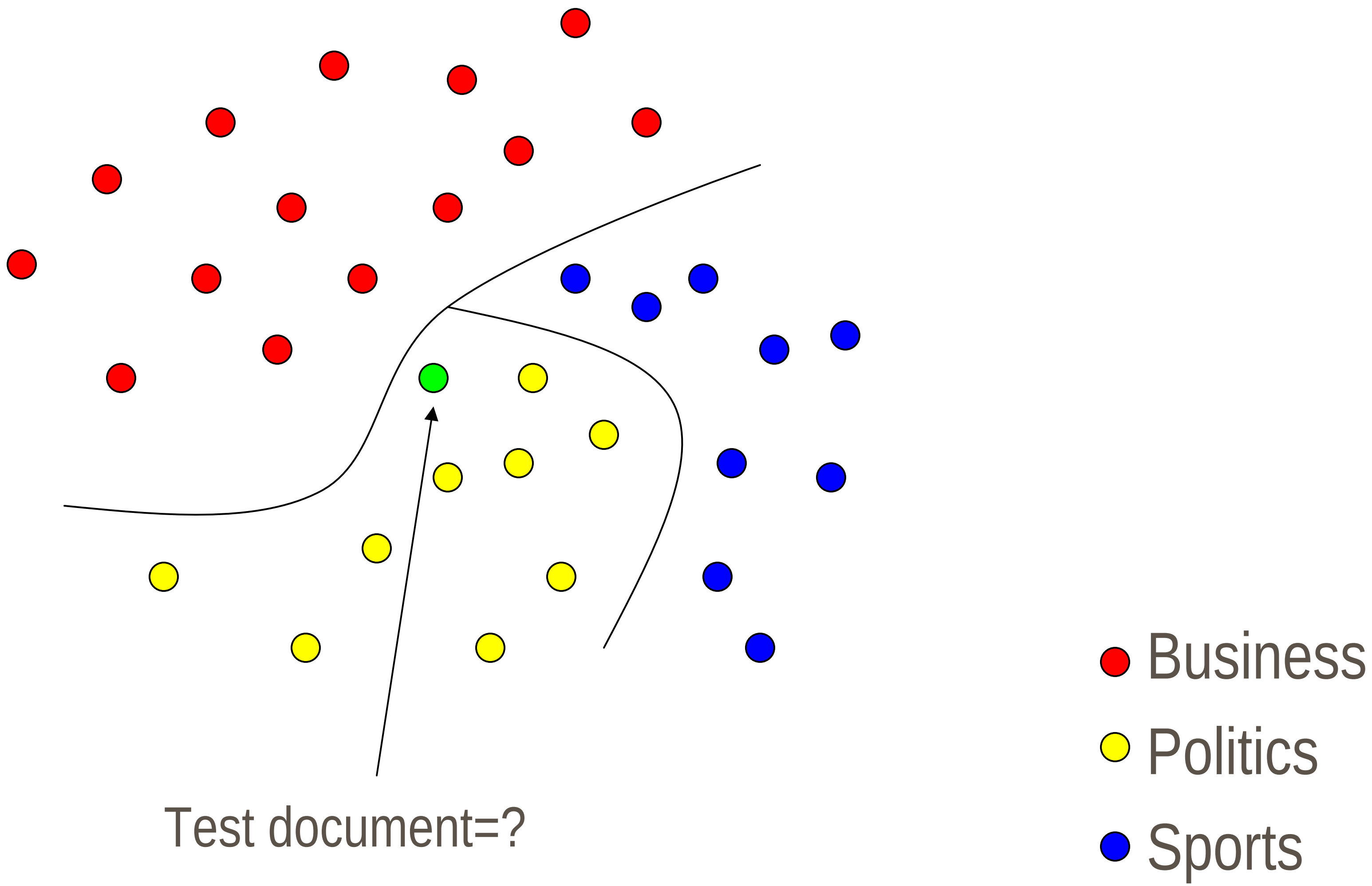
Text Classification

Setup

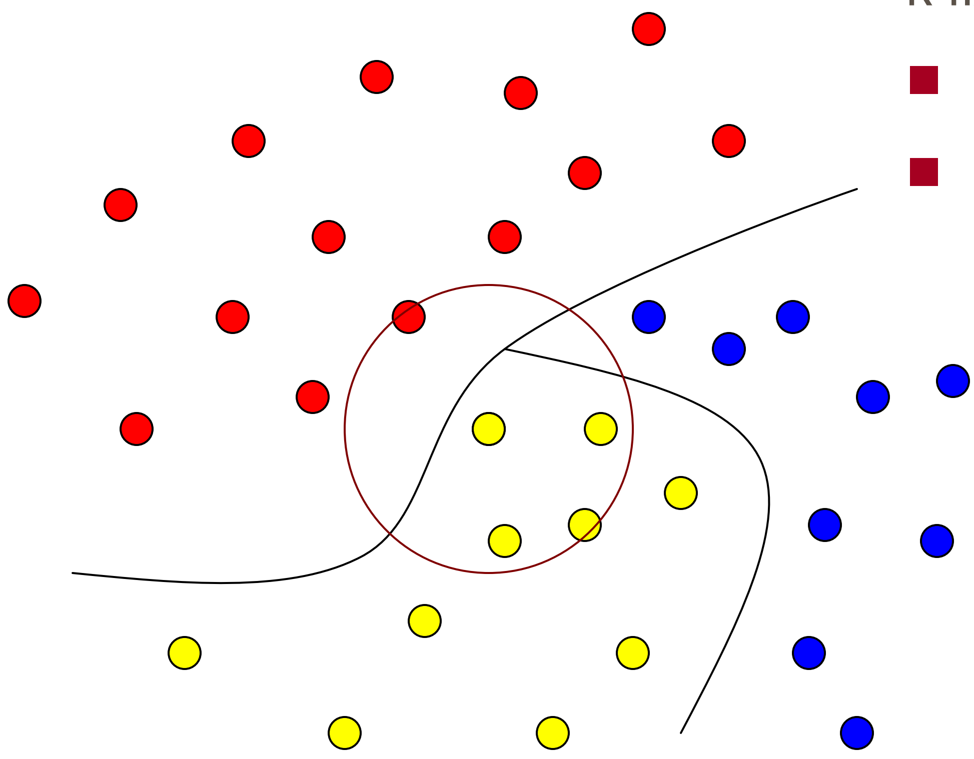
- Vocabulary (e.g. 35000 words)
- Document=vector of word occurrences $\in R^{35000}$
- Normalized to unit length
- Thousands of points for each class

Word	Doc1	Doc2	Doc3	…
abyss	0	3	2	
…				
budget	5	2	3	
…				
Clinton	6	4	1	
…				
Zaire	1	0	2	

Classes=Regions in Vector Space

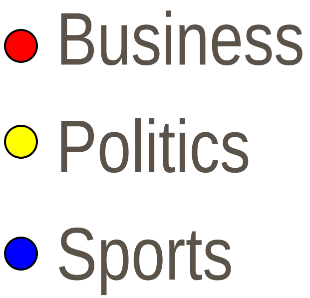


K-Nearest Neighbor (kNN) Classifier



k fixed (e.g. k=3)

- Find the k nearest neighbors
- Take the majority class



Instance Based Learning

- kNN is an example of Instance Based Learning
- Instance Based Learning
  - Store many examples (instances)
  - Distance metric to the examples
  - Value of k
    - number of examples to make decision from
  - Weighting function (optional)

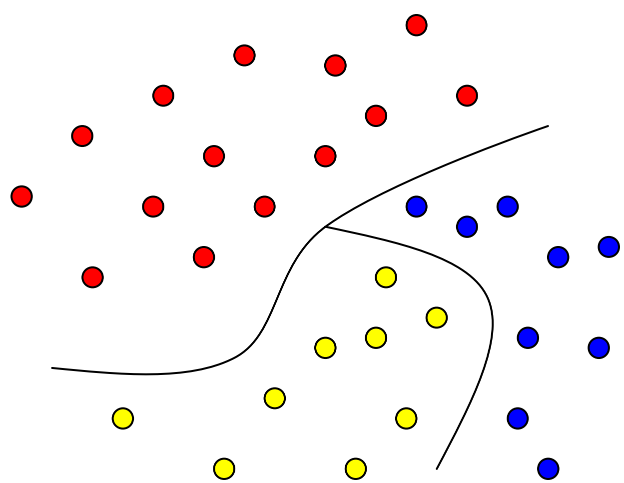
- Disadvantages
  - Classification is expensive (search problem)
  - Need to store many examples

Distance Metrics

- Euclidian Distance $d(x,x')^2 = \sum \sigma_i^2 (x_i - x'_i)^2$
- Mahalanobis Distance $d(x,x')^2 = (x-x')^T \Sigma (x-x')$
  - Σ is symmetric positive definite
- L_1 norm $d(x,x') = \sum |x_i - x'_i|$
- L_∞ norm $d(x,x') = \max |x_i - x'_i|$
- Correlation $d(x,x') = \dfrac{x^T x'}{|x||x'|}$
- Angle
- Hamming distance
- Manhattan distance
- …

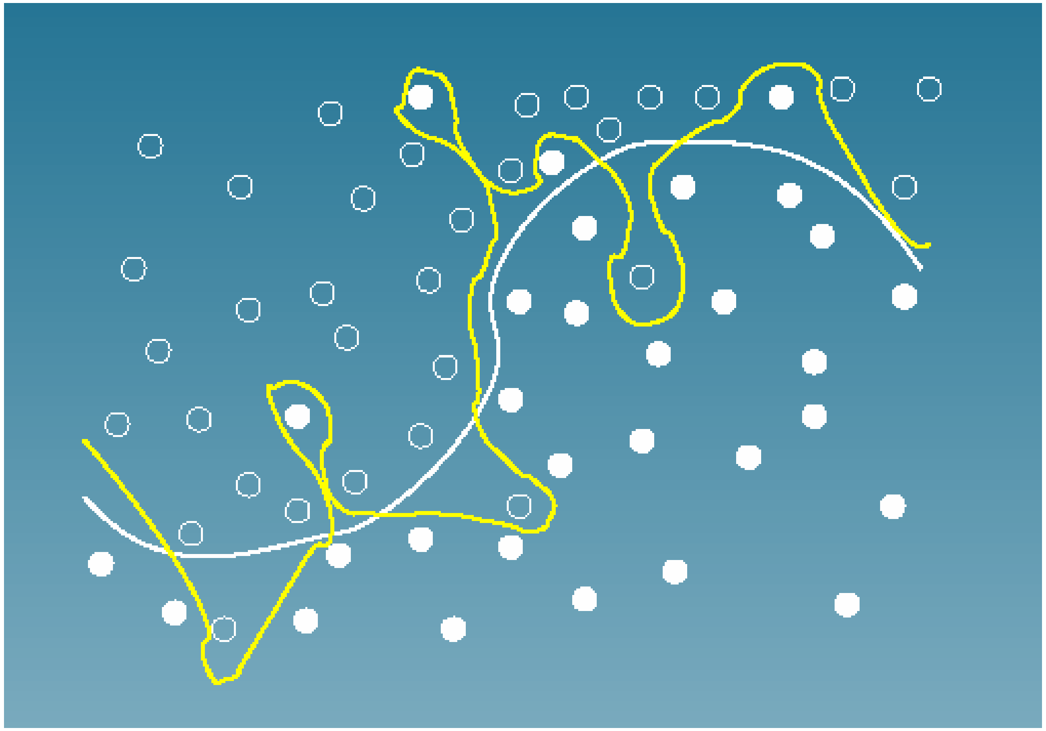
Optimality of kNN

- Cover and Hart 1967
- Bayes error rate
  - Error rate when you know the model that generated the data
  - Best you can do
- Asymptotically (when $N \to \infty$)
  - Error of 1-NN is less that 2*Bayes Error
  - In particular, Error of 1-NN $\to 0$ if the Bayes error is 0 i.e. classes are separable
  - Decision boundary



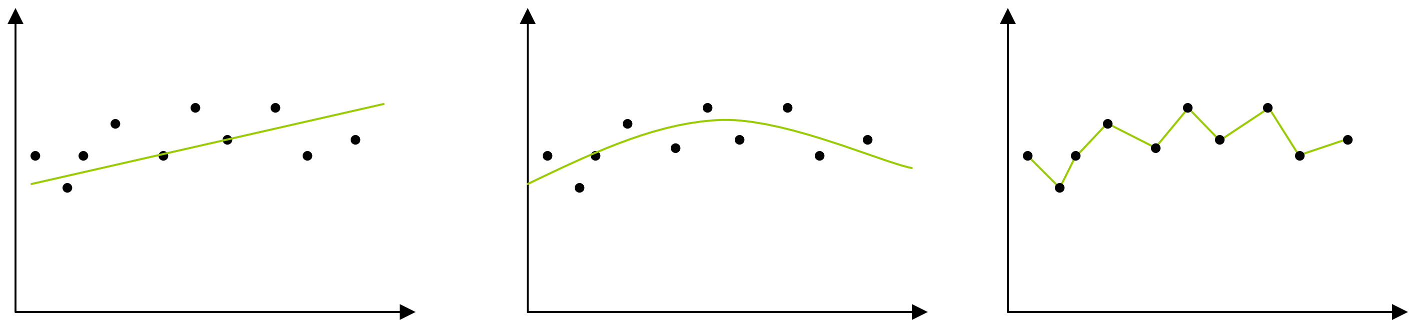
Overfitting

- In reality, classes are usually not separable
- Separating them → overfitting

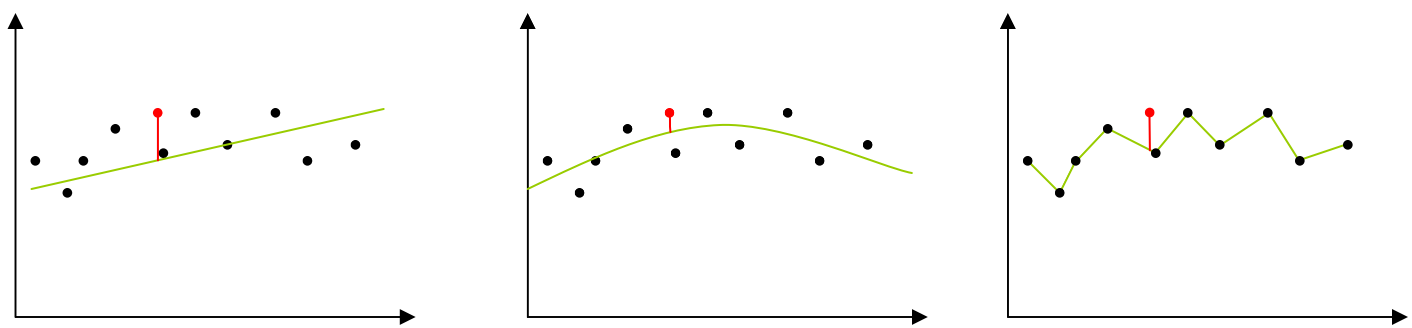


Overfitting for Regression

- Training



- Testing



Underfitting Overfitting

Bias-Variance Tradeoff

- Consider:
  - A training dataset D
  - A test sample x
  - A regression algorithm f trained on D gives f(x,D)
- Underlying truth
  - Given x, the output y comes from a probability P(y|x)
- Expected value $E[y|x] = \bar{y}$
  - E.g. say y∈ {1,2} and for a specific x, $P(y=1|x) = 0.1$ then $E[y|x] = 1 \cdot 0.1 + 2 \cdot 0.9 = 1.9$
- Measure of error for f:

$$E[(y - f(x,D))^2|x,D]$$

Bias-Variance Tradeoff

- Then

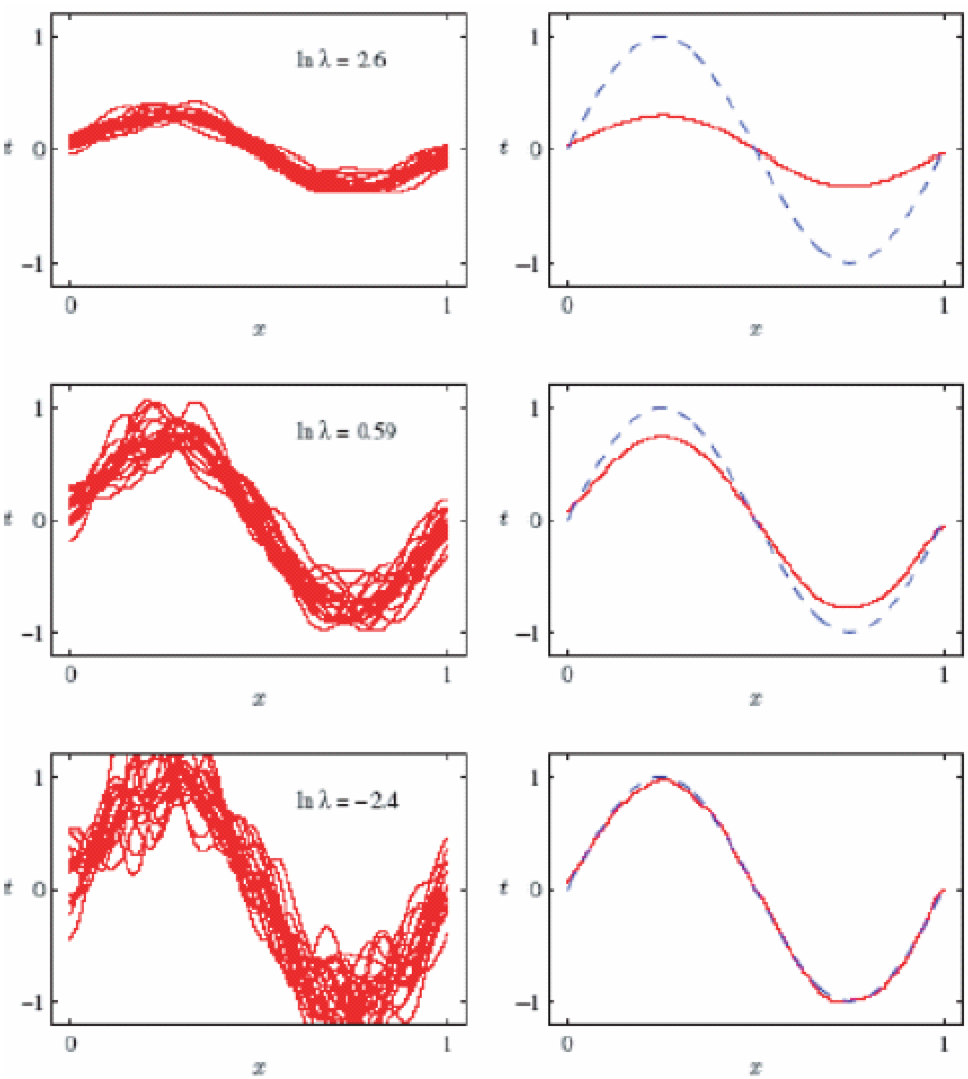
$$E[(y-f(x,D))^2|x,D] = E[(y-\bar{y})^2|x,D]+(f(x,D)-\bar{y})^2$$

- Variance
  - The term $E[(y-\bar{y})^2|x,D]$ is the variance of y, does not depend on D
- Now we look at the error over all training sets D

$$\begin{aligned} E_D[(f(x,D)-\bar{y})^2] = & (E_D[f(x,D)]-\bar{y})^2 & \text{bias} \\ & + E_D[(f(x,D)-E_D[f(x,D)])^2] & \text{variance} \end{aligned}$$

- Bias = how far is the average result from the avg. true result
- Variance= the variability of the result

Example



Results from 100 datasets

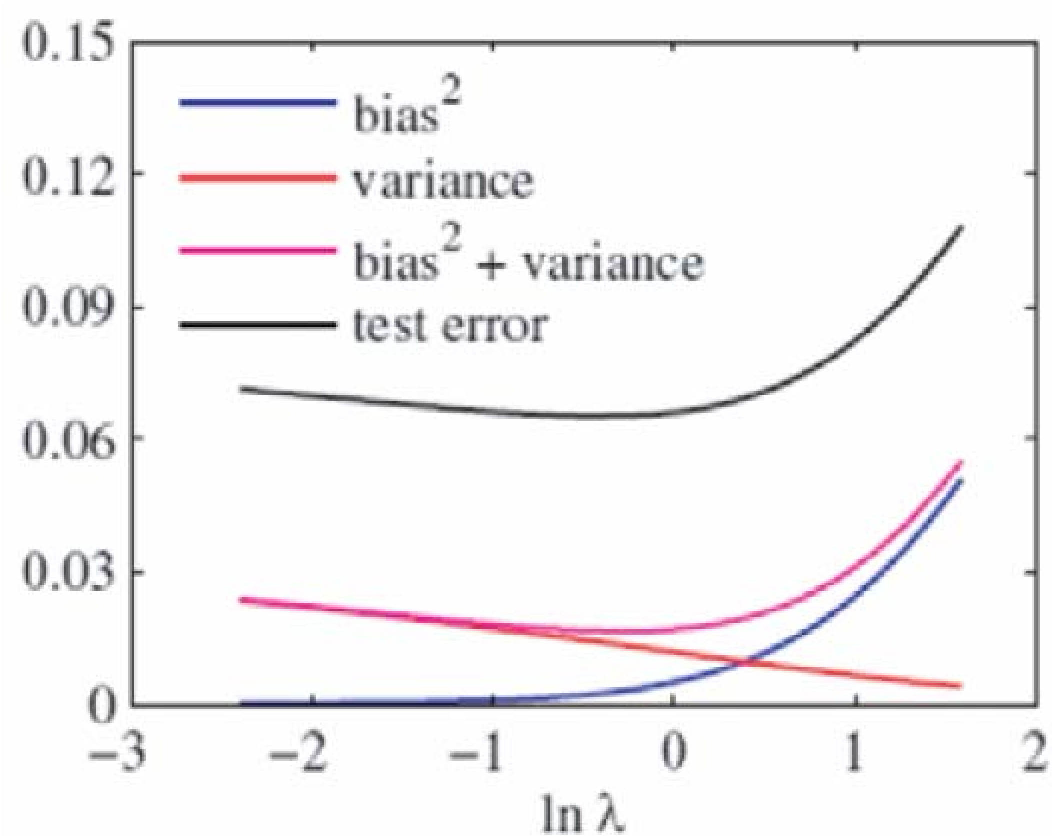
Average over 100 results

Regularized regression

$$l(W) = \frac{1}{N}\sum_{j=1}^{N}(X^j w - Y^j)^2 + \lambda \sum_{i=1}^{M} w_i^2$$

- Large bias small variance
- Small bias
  Large variance

Bias^2+Variance vs. λ



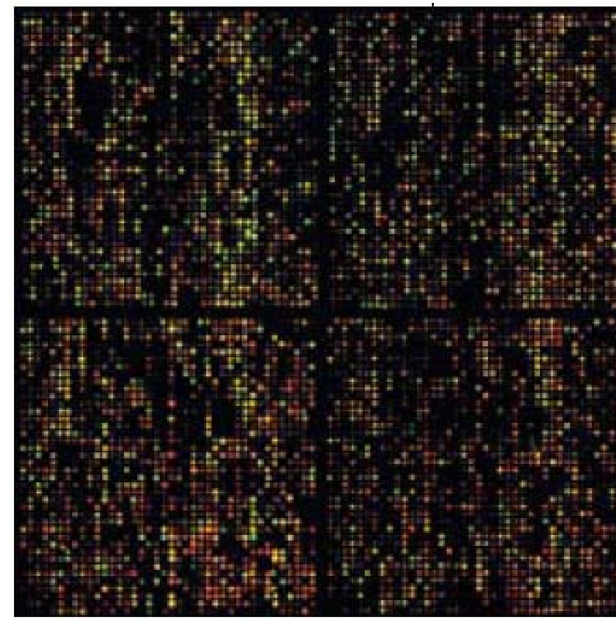
- Bias^2+variance has similar shape with test error
- However, bias and variance cannot be computed in general
  - We don't know the true distribution of X and Y

Feature Selection

- Regularization (Regularized Loss Functions)
- Wrappers:
  - Use cross-validation and any learning algorithm
  - Repeat for the desired number of features:
    - Add the feature that minimizes cross-validation error
  - Greedy and slow to train
- Feature Ranking (Xing et al, 2001)
  - Bayes Error
  - Information Gain
  - Markov Blanket
  - Faster

Feature Selection

Xing et al, 2001



- Dataset:
  - Expression levels for 7130 genes from a microarray dataset
  - 72 observations (samples)
  - 47 type I Leukemia (called ALL)
  - and 25 type II Leukemia (called AML)
- Need feature selection to avoid overfitting
- 3-stage feature selection
  - Mixture overlap probability
  - Information gain
  - Markov blanket

Feature Selection

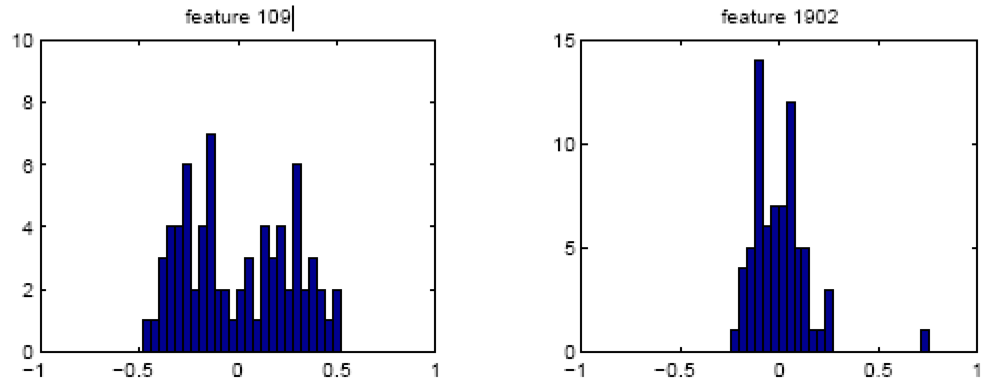
Xing et al, 2001

Features =

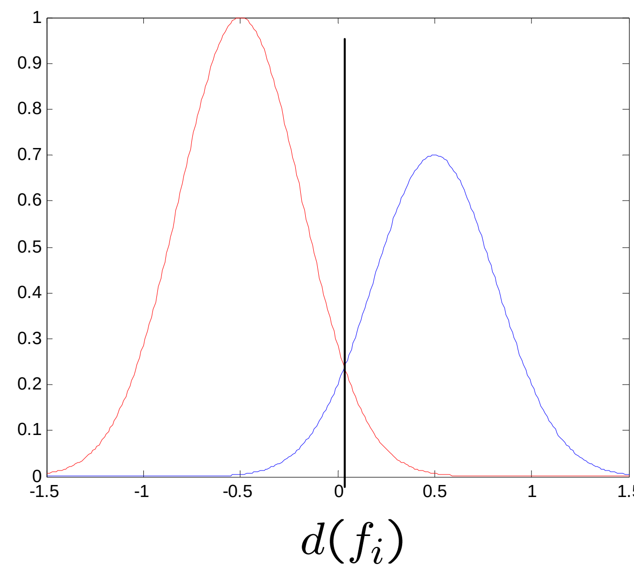
Gene expression levels

- Two hidden gene states:
  - Active or inactive
  - Denote by z_i
  - Mixture of two gaussians

    Fitted with EM



Histograms of gene expression levels



Stage 1: Mixture overlap probability

$$\epsilon = P(z_i = 0)P(d(f_i) = 1 | z_i = 0) + P(z_i = 1)P(d(f_i) = 0 | z_i = 1)$$

- Area of overlap of the two Gaussians
- Chooses features for which is clear when they are expressed

Feature Selection

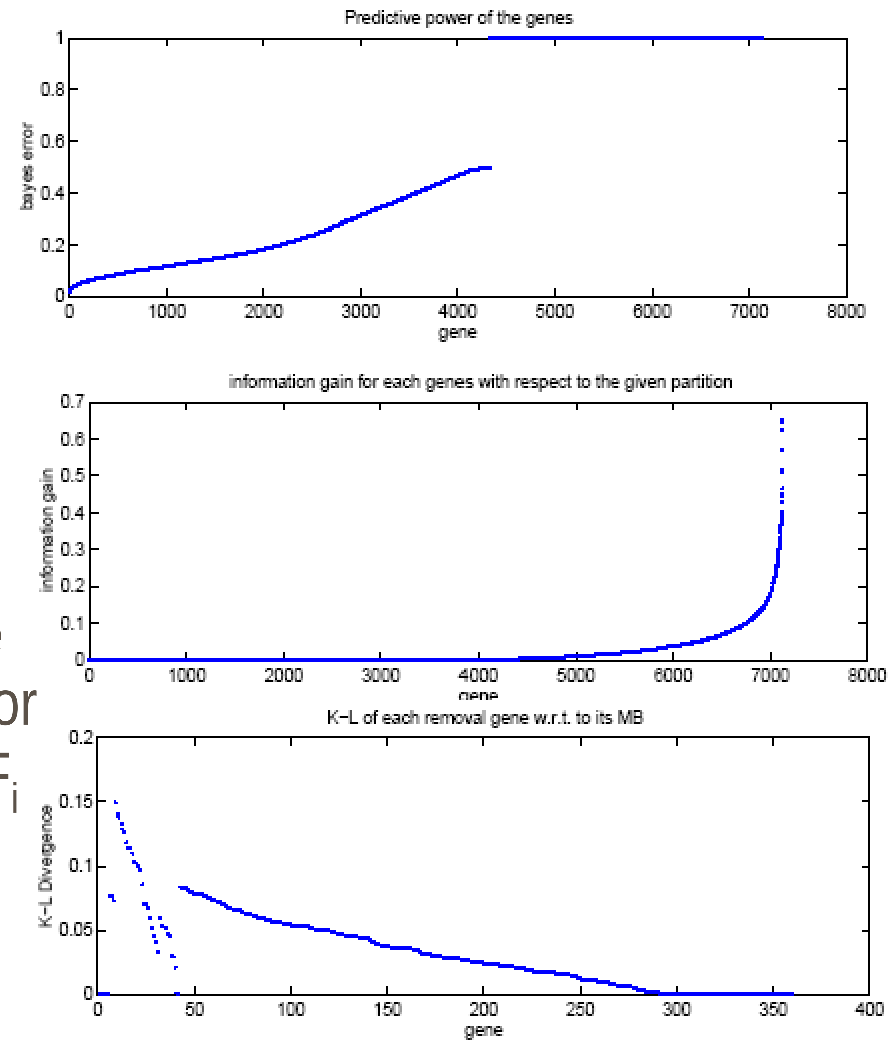
Xing et al, 2001

Stage 2: Information Gain

- Same as in decision trees
- Threshold from Stage 1
- Keep best 360 features

Stage 3: Markov Blanket Filtering

- Initialize G = F
- Iterate
  - For each feature $F_i \in G$, let M_i be the set of *k* features $F_j \in G-\{F_i\}$ for which the correlations between F_i and F_j are the highest.
  - Compute $\Delta(F_i|M_i)$ for each i
  - Choose the i that minimizes $\Delta(F_i|M_i)$, and define $G=G-\{F_i\}$

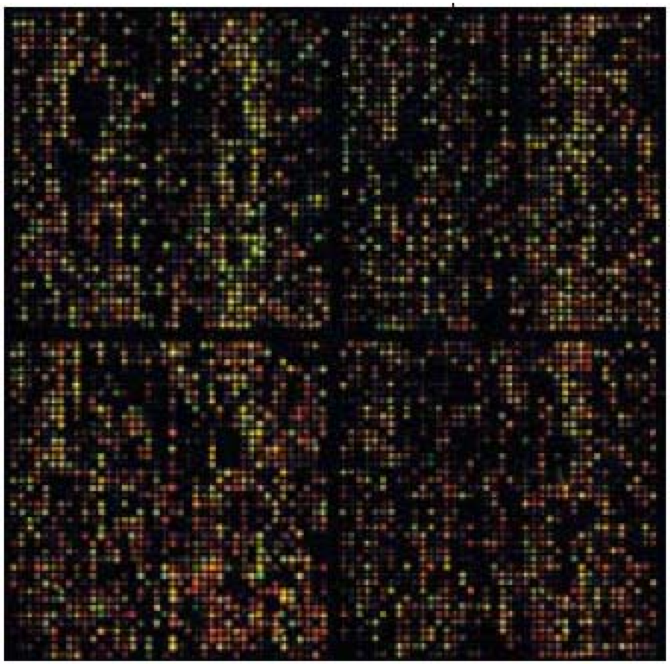


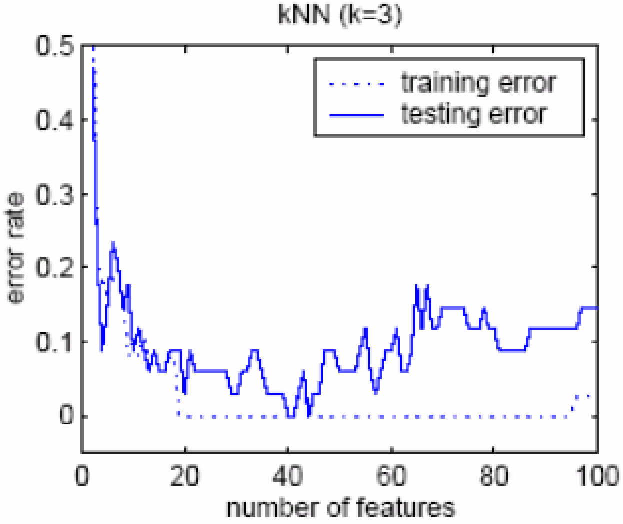
$$\Delta(F_i|M) = \sum_{f_M, f_i} P(M=f_M; F_i=f_i) D(P(C|M=f_M, F_i=f_i)||P(C|M=f_M)$$

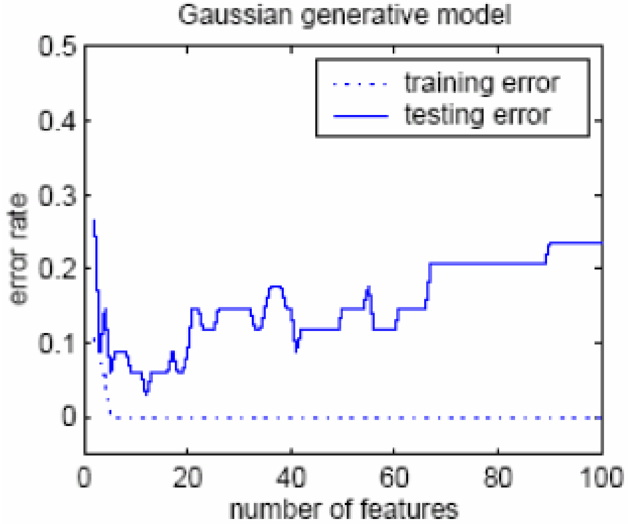
Feature Selection

Xing et al, 2001

- Obtain about 40 good features
- Learning algorithms:
  - kNN
  - Naïve Bayes with Gaussian models
  - Logistic regression





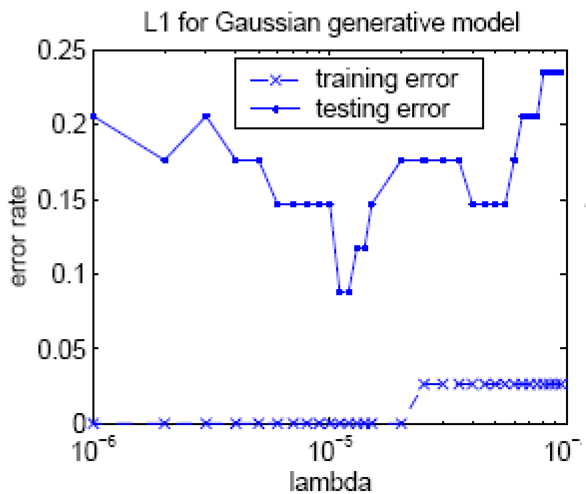


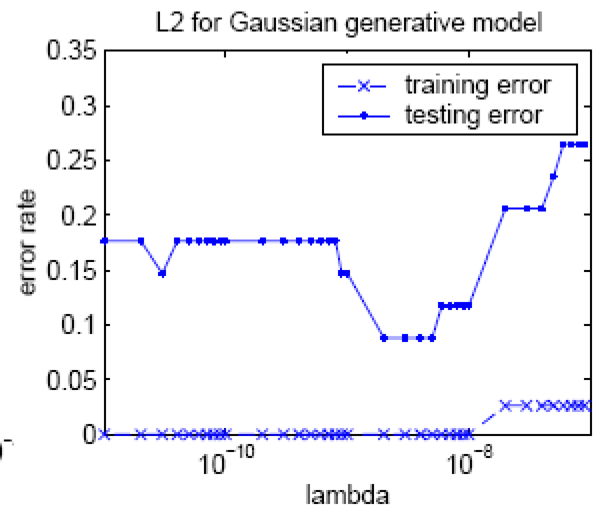
Error when gradually adding less qualified features

Feature Selection vs. Regularized Loss

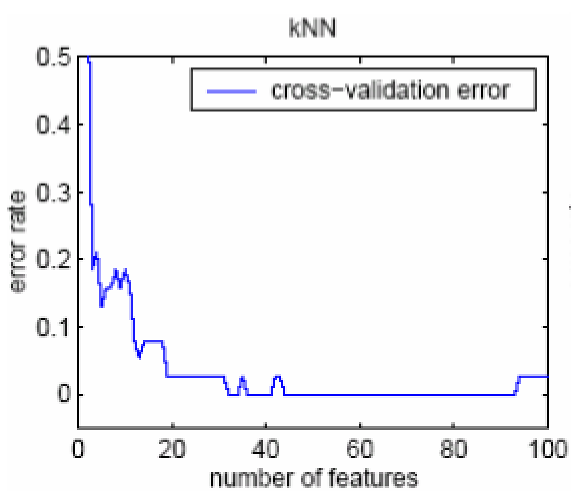
Feature selection outperforms Regularized Loss (regularization)

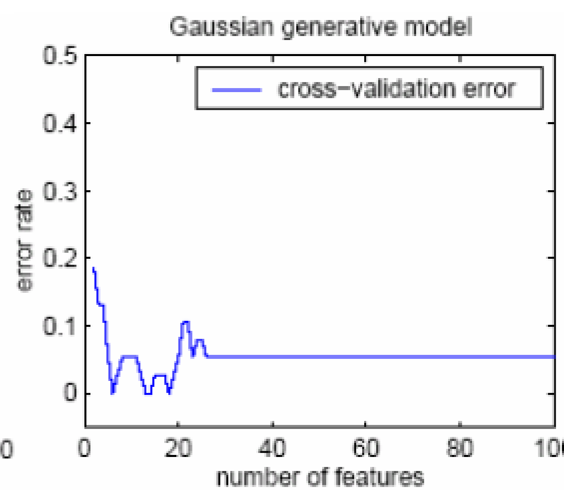
Regularization

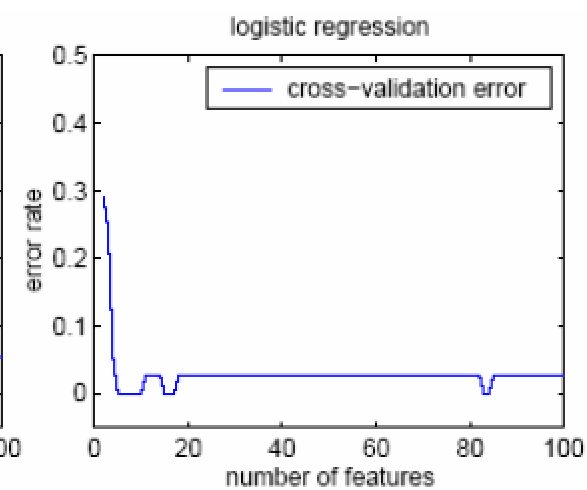




Feature Selection







References

- EP Xing, MI Jordan, RM Karp. Feature selection for high-dimensional genomic microarray data. ICML, 2001
- ASU Feature Selection Website:

  http://featureselection.asu.edu/index.php