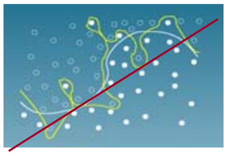
Machine Learning

10-701/15-781, Spring 2008

Model/Feature Selection

Eric Xing



Lecture 14, March 3, 2008

Reading: Chap. 1&2, CB & Chap 5,6, TM

Bias-variance decomposition

- For one data set D and one test point x

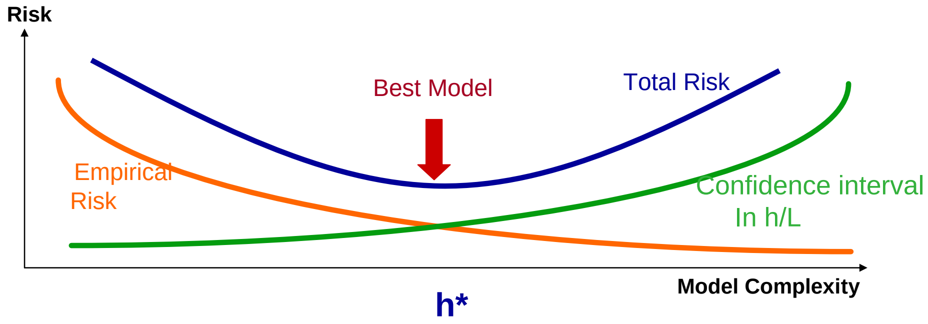
$$
\begin{aligned}
& E_{(x,t),D}\left[(y(x)-t)^2\right] \\
&= \int \left(E_D\left[y(x;D)\right]-h(x)\right)^2 p(x)dx \\
&\quad + \int E_D\left[\left(y(x;D)-E_D\left[y(x;D)\right]\right)^2\right] p(x)dx \\
&\quad + \int \left(h(x)-t\right)^2 p(x,t)dxdt
\end{aligned}
$$

$\Rightarrow$ expected loss = (bias)^2 + variance + noise

- Recall the VC bound:

$$\epsilon(h) \leq \hat{\epsilon}(h) + O\left(\sqrt{\frac{d}{m}\log\frac{m}{d} - \frac{1}{m}\log\delta}\right)$$

Minimizing
Empirical Risk & Structural Risk



SRM & ERM in practice

- There are many SRM-based strategies to build models:

- In the case of linear models

  y = <w|x> + b,

  one wants to make ||w|| a controlled parameter: let us call H_C the linear model function family satisfying the constraint:

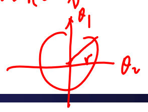
  ||w|| < C

> Vapnik Major theorem:
>
> When C decreases, $d(H_C)$ decreases
>
> ||x|| < R

Regularization

$\theta = \begin{pmatrix}\theta_1 \\ \theta_2\end{pmatrix}$

$\|\theta\|_2 = \sqrt{\theta_1^2 + \theta_2^2}$



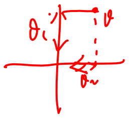
- Maximum-likelihood estimates are not always the best (James and Stein showed a counter example in the early 60's)
- Alternative: we "regularize" the likelihood objective (also known as penalized likelihood, shrinkage, smoothing, etc.), by adding to it a penalty term:

$\|\theta\|_1 = |\theta_1| + |\theta_2|$

$$\hat{\theta}_{\text{shrinkage}} = \arg\max_{\theta}\left[l(\theta; D) + \lambda\|\theta\|\right]$$

where λ>0 and $||\theta||$ might be the L_1 or L_2 norm.

$\|\theta\|_\infty$



- The choice of norm has an effect
  - using the L_2 norm pulls directly towards the origin,
  - while using the L_1 norm pulls towards the coordinate axes, i.e it tries to set some of the coordinates to 0.
  - This second approach can be useful in a feature-selection setting.

Bayesian and Frequentist

- Frequentist interpretation of probability
  - Probabilities are objective properties of the real world, and refer to limiting relative frequencies (e.g., number of times I have observed heads). Hence one cannot write *P*(Katrina could have been prevented|*D*), since the event will never repeat.
  - Parameters of models are *fixed, unknown constants*. Hence one cannot write *P*(*θ*|*D*) since *θ* does not have a probability distribution. Instead one can only write *P*(*D*|*θ*).
  - One computes point estimates of parameters using various *estimators*, *θ**= *f*(*D*), which are designed to have various desirable qualities when *averaged over future data D* (assumed to be drawn from the "true" distribution).
- Bayesian interpretation of probability
  - Probability describes degrees of belief, not limiting frequencies.
  - Parameters of models are *hidden variables*, so one can compute *P*(*θ*|*D*) or *P*(*f*(*θ*)|*D*) for some function *f*.
  - One estimates parameters by computing *P*(*θ*|*D*) using Bayes rule:

$$p(\theta|D) = \frac{p(D\,|\,\theta)\,p(\theta)}{p(D)}$$

$P(\theta)$

$\Rightarrow \int f(\theta)\, p(\theta)\, d\theta$

$\searrow$

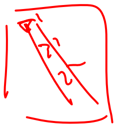
$\hat{\theta}_{Bayes} = \int \theta\, p(\theta)\, d\theta$

Bayesian interpretation of regulation

- Regularized Linear Regression
  - Recall that using squared error as the cost function results in the LMS estimate
  - And assume iid data and Gaussian noise, LMS is equivalent to MLE of θ

$$l(\theta) = n \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2}\frac{1}{2}\sum_{i=1}^{n}(y_i - \theta^T \mathbf{x}_i)^2$$

  - Now assume that vector θ follows a normal prior with 0-mean and a diagonal covariance matrix



$$\theta \sim N(0, \tau^2 I)$$

  - What is the posterior distribution of θ?

$$p(\theta|D) \propto p(D,\theta)$$

$$= p(D|\theta)\,p(\theta) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_n - \theta^T x_i\right)^2\right\} \times C \exp\left\{-(\theta^T\theta/2\tau^2\right\}$$

Bayesian interpretation of regulation, con'd

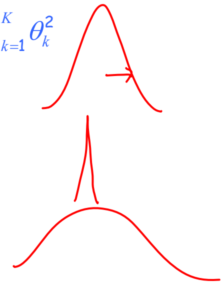
- The posterior distribution of θ

$$p(\theta|D) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_n - \theta^T x_i\right)^2\right\} \times \exp\left\{-\theta^T\theta/2\tau^2\right\}$$

- This leads to a now objective

$$l_{MAP}(\theta; D) = -\frac{1}{2\sigma^2}\frac{1}{2}\sum_{i=1}^{n}(y_i - \theta^T \mathbf{x}_i)^2 - \frac{1}{\tau^2}\frac{1}{2}\sum_{k=1}^{K}\theta_k^2$$

$$= l(\theta; D) + \lambda\|\theta\|$$



  - This is L_2 regularized LR! --- a MAP estimation of θ
  - What about L_1 regularized LR! (homework)
- How to choose λ.
  - cross-validation!

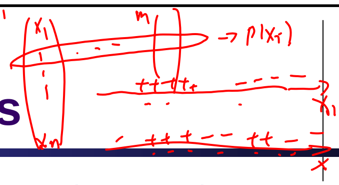
Feature Selection

$\theta_i \to 0$

$f(\theta^T x)$

- Imagine that you have a supervised learning problem where the number of features n is very large (perhaps n >>#samples), but you suspect that there is only a small number of features that are "**relevant**" to the learning task.
- VC-theory can tell you that this scenario is likely to lead to high generalization error – the learned model will potentially overfit unless the training set is fairly large.
- So lets get rid of useless parameters!

How to score features

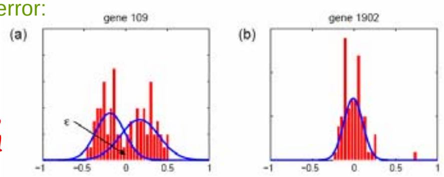


- How do you know which features can be pruned?
  - Given labeled data, we can compute some simple score $S(i)$ that measures how informative each feature x_i is about the class labels y.
  - Ranking criteria:
    - Mutual Information: score each feature by its mutual information with respect to the class labels

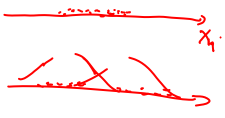
$$MI(x_i, y) = \sum_{x_i \in \{0,1\}} \sum_{y \in \{0,1\}} p(x_i, y) \log \frac{p(x_i, y)}{p(x_i)p(y)}$$

$x_i^{(1)}$ $x_i^{(2)}$ ... $x_i^{(m)}$

    - Bayes error:



mixture of G

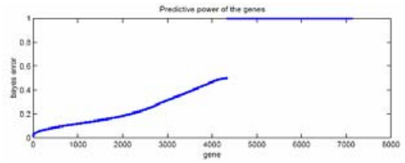


x_i x_j x_k

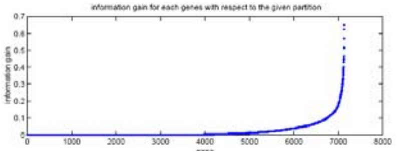
    - Redundancy (Markov-blank score) …
  - We need estimate the relevant p()'s from data, e.g., using MLE

Feature Ranking

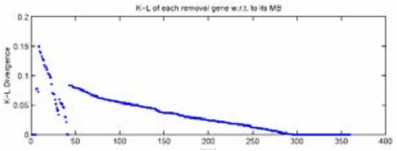
- Bayes error of each gene



- information gain for each genes with respect to the given partition



- KL of each removal gene w.r.t. to its MB



Feature selection schemes

- Given n features, there are 2^n possible feature subsets (why?)
- Thus feature selection can be posed as a model selection problem over 2^n possible models.
- For large values of n, it's usually too expensive to explicitly enumerate over and compare all 2^n models. Some heuristic search procedure is used to find a good feature subset.
- Three general approaches:
  - Filter: i.e., direct feature ranking, but taking no consideration of the subsequent learning algorithm
    - add (from empty set) or remove (from the full set) features one by one based on $S(i)$
    - Cheap, but is subject to local optimality and may be unrobust under different classifiers
  - Wrapper: determine the (inclusion or removal of) features based on performance under the learning algorithms to be used. See next slide
  - Simultaneous learning and feature selection.
    - E.x. L_1 regularized LR, Bayesian feature selection (will not cover in this class), etc.

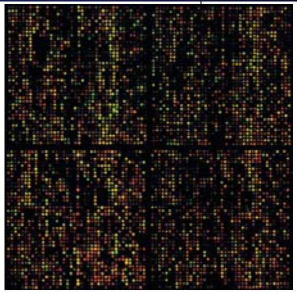
Wrapper

- Forward:
  1. Initialize $\mathcal{F} = \varnothing$
  2. Repeat
     - For $i = 1, \dots, n$

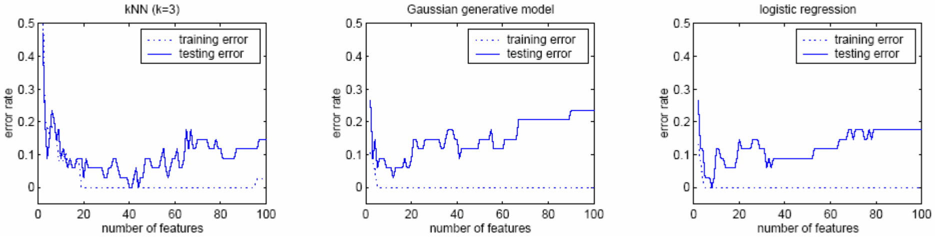
       if $i \notin \mathcal{F}$, let $\mathcal{F}_i = \mathcal{F} \cup \{i\}$, and use some version of cross validation to evaluate features $\mathcal{F}_i$. (I.e., train your learning algorithm using only the features in $\mathcal{F}_i$, and estimate its generalization error.)
     - Set $\mathcal{F}$ to be the best feature subset found on the last step step.
  3. Select and output the best feature subset that was evaluated during the entire search procedure.

- Backward search
  1. Initialize $\mathcal{F}$ = full set
  2. …

Case study [Xing et al, 2001]

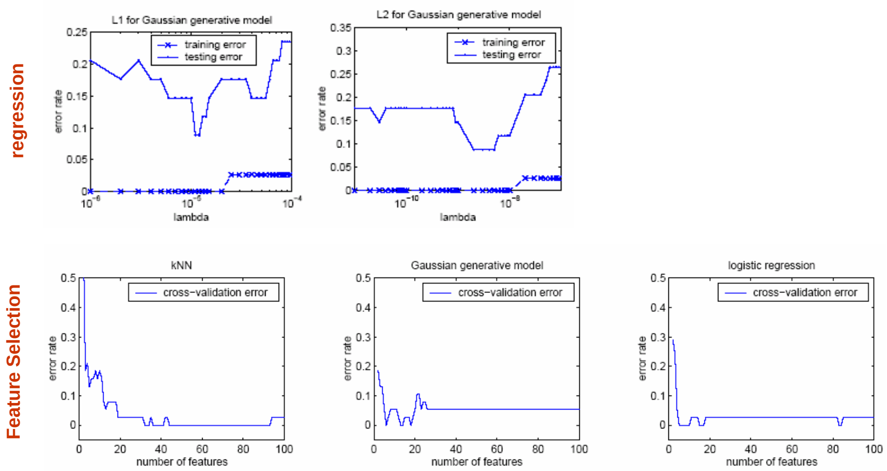


- The case:
  - **7130 genes from a microarray dataset**
  - **72 samples**
  - **47 type I Leukemias (called ALL) and 25 type II Leukemias (called AML)**
- Three classifier:
  - **kNN**
  - **Gaussian classifier**
  - **Logistic regression**



Regularization vs. Feature Selection

- Explicit feature selection often outperform regularization



Model Selection

- Suppose we are trying select among several different models for a learning problem.
- Examples:

  1. polynomial regression

  $$h(x;\theta) = g(\theta_0 + \theta_1 x + \theta_2 x^2 + \ldots + \theta_k x^k)$$

  - Model selection: we wish to **automatically** and **objectively** decide if k should be, say, 0, 1, . . . , or 10.

  2. locally weighted regression,

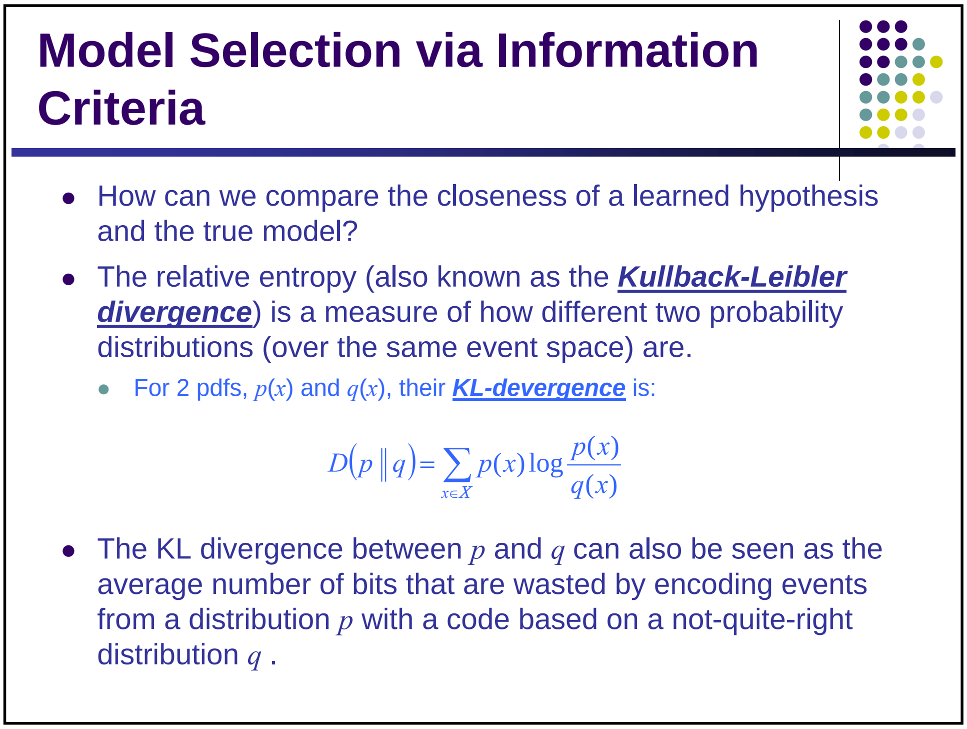
  - Model selection: we want to automatically choose the bandwidth parameter τ.

  3. Mixture models and hidden Markov model,

  - Model selection: we want to decide the number of hidden states

- The Problem:
  - Given model family $\mathcal{F} = \{M_1, M_2, \ldots, M_I\}$, find $M_i \in \mathcal{F}$ s.t.

  $$M_i = \arg\max_{M \in \mathcal{F}} J(D, M)$$

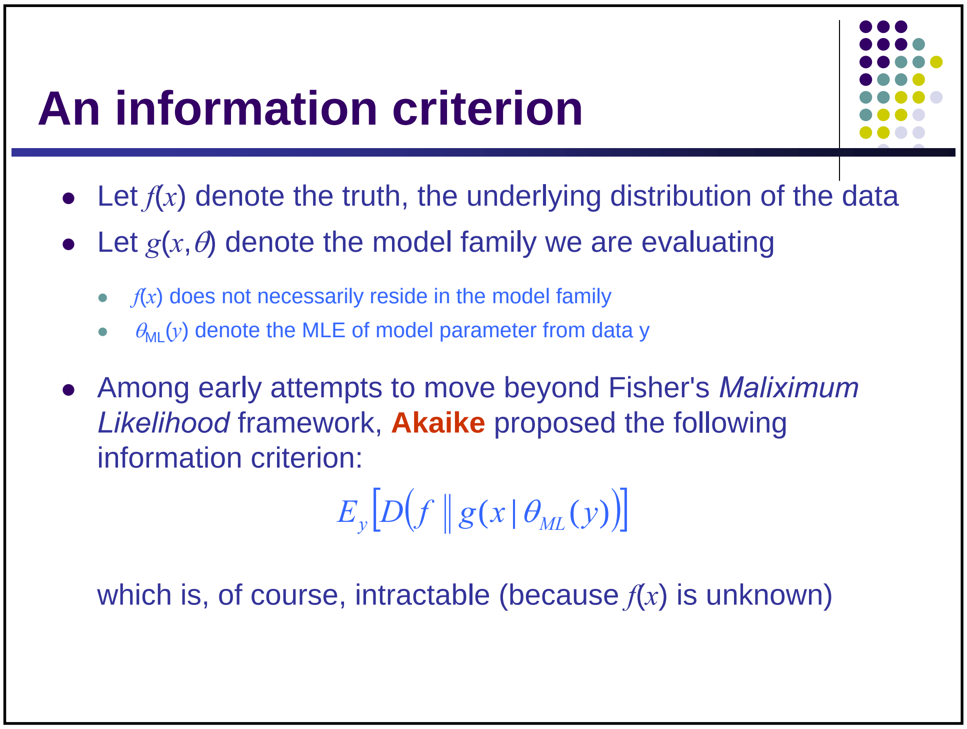


Model Selection via Information Criteria

- How can we compare the closeness of a learned hypothesis and the true model?
- The relative entropy (also known as the ***Kullback-Leibler divergence***) is a measure of how different two probability distributions (over the same event space) are.
  - For 2 pdfs, $p(x)$ and $q(x)$, their ***KL-devergence*** is:

$$D(p \parallel q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

- The KL divergence between p and q can also be seen as the average number of bits that are wasted by encoding events from a distribution p with a code based on a not-quite-right distribution q .



An information criterion

- Let $f(x)$ denote the truth, the underlying distribution of the data
- Let $g(x,\theta)$ denote the model family we are evaluating
  - $f(x)$ does not necessarily reside in the model family
  - $\theta_{ML}(y)$ denote the MLE of model parameter from data y
- Among early attempts to move beyond Fisher's *Maliximum Likelihood* framework, **Akaike** proposed the following information criterion:

$$E_y\left[D\left(f \parallel g(x \mid \theta_{ML}(y))\right)\right]$$

which is, of course, intractable (because $f(x)$ is unknown)

AIC and TIC

- AIC (A information criterion, not **Akaike** information criterion)

$$A = \log g(x \mid \hat{\theta}(y)) - k$$

where k is the number of parameters in the model

- TIC (Takeuchi information criterion)

$$A = \log g(x \mid \hat{\theta}(y)) - \mathrm{tr}(I(\theta_0)\Sigma)$$

where

$$\theta_0 = \arg\min D(f \parallel g(\cdot \mid \theta)) \qquad I(\theta_0) = -E_x\left[\frac{\partial^2 \log g(x \mid \theta)}{\partial\theta\partial\theta^T}\right]\Bigg|_{\theta=\theta_0} \qquad \Sigma = E_y\left(\hat{\theta}(y) - \theta_0\right)\left(\hat{\theta}(y) - \theta_0\right)^T$$

  - We can approximate these terms in various ways (e.g., using the bootstrap)
  - $\mathrm{tr}(I(\theta_0)\Sigma) \approx k$

Bayesian Model Selection

- Recall the Bayesian Theory: (e.g., for date *D* and model *M*)

$$P(M|D) = P(D|M)P(M)/P(D)$$

  - the **posterior** equals to the **likelihood** times the **prior**, up to a constant.

- Assume that $P(M)$ is uniform and notice that $P(D)$ is constant, we have the following criteria:

$$P(D \mid M) = \int_\theta P(D \mid \theta, M) P(\theta \mid M) d\theta$$

- A few steps of approximations (you will see this in advanced ML class in later semesters) give you this:

$$P(D \mid M) \approx \log P(D \mid \hat{\theta}_{ML}) - \frac{k}{2} \log N$$

where N is the number of data points in D.

Summary

- Bias-variance decomposition
- The battle against overfitting:
    - Cross validation
    - Regularization
    - Model selection --- Occam's razor
    - Model averaging
        - The Bayesian-frequentist debate
        - Bayesian learning (weight models by their posterior probabilities)

Review

Method	Input.	Output	Loss	hypothesis	Opt procedure	~~generative~~ Discriminative
Density Est.	$X \in R^n$ $X \in D^n$.	$p(x)$	$L(\theta)$,	Gaussian. Multi. Parzen.	take deri. $=0$ → close form. → Gradient	—
linear Reg	$X \in R^n$ $y \in R$	$y = f(x)$ $= f(\theta^T x)$	$(\hat{f}(x) - y)^2$ $L(\theta)$	linear in feature x^2 x^1	Normal Eq Gradient: $\nearrow$ $\searrow$ (1) stochastic (2) steepest	
kNN	$X \in \underline{R}^n$ $y \in C$	$y \leftarrow f(x)$		<u>Parzen</u>		
NB.	"	$P(x\|y)$ $P(y)$ → $P(y\|x)$	$L(\theta)$,	linear. $(\Sigma_1 \equiv \Sigma_2)$ 2nd order o/w	<u>MLE</u>, <u>MAP</u>. Grad. close-form	G.

Review

Learning Theory : Bayes Opt Classifier

PAC

Agnostic — finite / infinite → VC.

Method	Input.	Output	Loss	hypothesis	Opt procedure	~~generative~~ Discriminative
Logistic Reg		$f(x) \to$ θ.	sq.		Gradient, – online / batch (LMS)	D.
ANN.	$X \in R^n$ $y \in R^m$	$\vec{W}$. for every perceptron.	sq	anything	back-prop (recursive two-pass) hidden variable,	D
SVM	$X \in R^n$ $y \in \mathbb{C}$.	$f(x) = w \cdot x + b$ $W = \sum_{i \in SV} \alpha_i X_i$	Margin	$\phi(x)$ in Hilbert Space $K(\cdot,\cdot)$ in mercer	QP. Dual–Primal Convex Opt.	D.
boosting						