# Score Function for Data Mining Algorithms

Chapter 7 of HTF

David Madigan

# Algorithm Components

1. The *task* the algorithm is used to address (e.g. classification, clustering, etc.)

2. The *structure* of the model or pattern we are fitting to the data (e.g. a linear regression model)

3. The *score function* used to judge the quality of the fitted models or patterns (e.g. accuracy, BIC, etc.)

4. The *search or optimization method* used to search over parameters and/or structures (e.g. steepest descent, MCMC, etc.)

5. The *data management technique* used for storing, indexing, and retrieving data (critical when data too large to reside in memory)

# Introduction

- e.g. how to pick the "best" $a$ and $b$ in $Y = aX + b$
- usual score in this case is the sum of squared errors
- Scores for patterns versus scores for models
- Predictive versus descriptive scores
- Typical scores are a poor substitute for utility-based scores

# Predictive Model Scores

$$S_{\mathrm{SSE}}(\theta) = \frac{1}{N}\sum_{i=1}^{N}(\hat{f}(x_i;\theta) - y_i)^2$$

$$S_{0/1}(\theta) = \frac{1}{N}\sum_{i=1}^{N} I(\hat{f}(x_i;\theta), y_i)$$

- Assume all observations equally important
- Depend on differences rather than values
- Symmetric
- Proper scoring rules

# Probabilistic Model Scores

$$L(\theta) = \prod_{i=1}^{N} \hat{p}(x_i);\theta)$$

- "Pick the model that assigns highest probability to what actually happened"
- Typically evaluated at the MLE

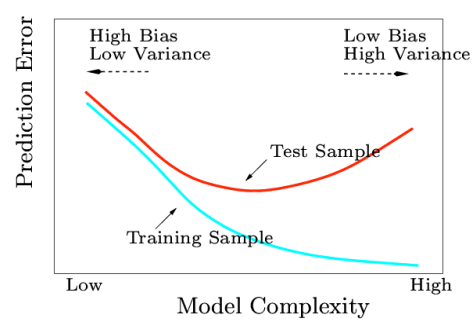
# Optimism of the Training Error Rate

- Typically the training error rate:

$$\overline{\text{err}} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$$

is an optimistically biased estimate of the true error rate:

$$\mathrm{E}_{X,y}\left[ L(y_i, \hat{f}(x_i)) \right]$$



# "in-sample" error

- Consider the error rate with fixed $x$'s

$$\mathrm{Err}_{\mathrm{in}} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{E}_y \mathrm{E}_{Y^{new}} L(Y_i^{new}, \hat{f}(x_i))$$

- Can show squared-error, 0-1, and other loss functions:

$$\mathrm{Err}_{\mathrm{in}} - \mathrm{E}_y(\overline{\mathrm{err}}) = \frac{2}{N}\sum_{i=1}^{N} Cov(\hat{y}_i, y_i)$$

- For the standard linear model with $p$ predictors:

$$\mathrm{Err}_{\mathrm{in}} = \mathrm{E}_y(\overline{\mathrm{err}}) + 2\frac{p}{N}\sigma_\varepsilon^2$$

# $C_p$ and AIC

•Leads directly to the $C_p$ statistic for scoring models

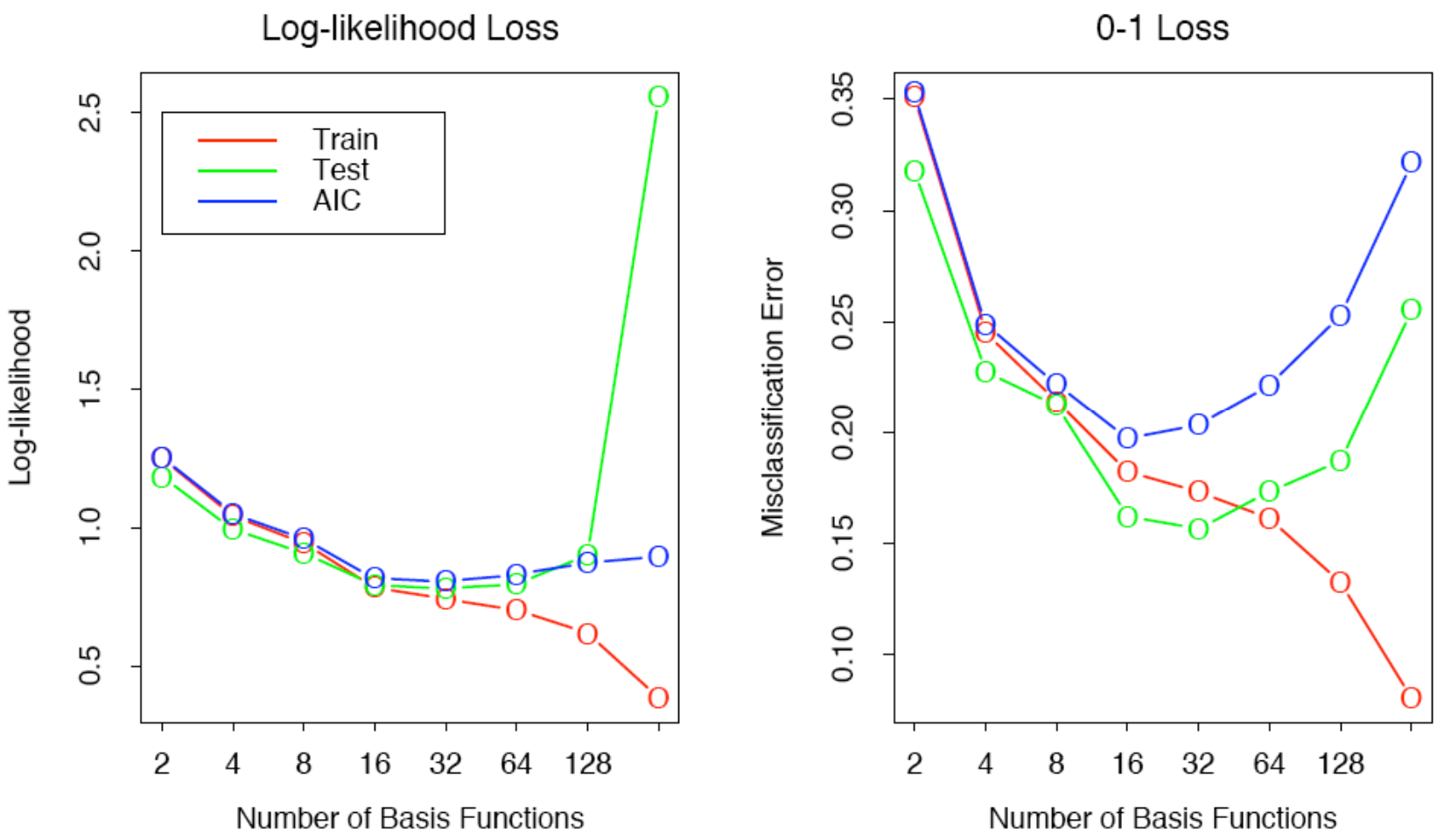
$$C_p = \overline{\text{err}} + 2\frac{p}{N}\hat{\sigma}_\varepsilon^2$$

•The Akaike Information Criterion (AIC) is derived simialrly but applies more generally to log likelihood loss

$$\text{AIC} = -\frac{2}{\text{N}} \cdot \text{loglik} + 2\frac{p}{N}$$

where loglik is the maximized log likelihood. AIC coincides with $C_p$ for the linear model

Selects overly complex models?


Log-likelihood Loss
Train
Test
AIC
Log-likelihood
0.5
1.0
1.5
2.0
2.5
2
4
8
16
32
64
128
Number of Basis Functions
0-1 Loss
Misclassification Error
0.10
0.15
0.20
0.25
0.30
0.35
2
4
8
16
32
64
128
Number of Basis Functions

# Bayesian Criterion

$$p(M_k \mid D) \propto p(D \mid M_k)\, p(M_k)$$

$$= p(M_k) \int p(D \mid \theta_k, M_k)\, p(\theta_k \mid M_k)\, d\theta_k$$

- Typically impossible to compute analytically
- All sorts of approximations

# Laplace Method for $p(D|M)$

$$\text{let } \; l(\theta) = \frac{\log(L(\theta))}{n} + \frac{\log p(\theta)}{n}$$

(i.e., the log of the integrand divided by $n$)

$$\text{then } \; p(D) = \int e^{nl(\theta)} d\theta$$

Laplace's Method:

$$p(D) \approx \int \exp[nl(\tilde{\theta}) - n(\theta - \tilde{\theta})^2 / (2\sigma^2)] d\theta$$

where $\sigma^2 = -1/l''(\tilde{\theta})$ and

$\tilde{\theta}$ is the posterior mode

# Laplace cont.

$$p(D) = \int \exp[nl(\tilde{\theta}) - n(\theta - \tilde{\theta})^2 / (2\sigma^2)] d\theta$$

$$\approx \sqrt{2\pi}\sigma n^{-1/2} \exp\{nl(\tilde{\theta})\}$$

- Tierney & Kadane (1986, JASA) show the approximation is $O(n^{-1})$
- Using the MLE instead of the posterior mode is also $O(n^{-1})$
- Using the expected information matrix in $\sigma$ is $O(n^{-1/2})$ but convenient since often computed by standard software
- Raftery (1993) suggested approximating $\tilde{\theta}$ by a single Newton step starting at the MLE
- Note the prior is explicit in these approximations

# Monte Carlo Estimates of $p(D|M)$

$$p(D) = \int p(D|\theta)\, p(\theta)\, d\theta$$

Draw iid $\theta_1, \ldots, \theta_m$ from $p(\theta)$:

$$\hat{p}(D) = \frac{1}{m} \sum_{i=1}^{m} p(D|\theta^{(i)})$$

In practice has large variance

# Monte Carlo Estimates of $p(D|M)$ (cont.)

Draw iid $\theta_1, \ldots, \theta_m$ from $p(\theta|D)$:

$$\hat{p}(D) = \frac{\frac{1}{m}\sum_{i=1}^{m} w_i \, p(D|\theta^{(i)})}{\sum_{i=1}^{m} w_i}$$

"Importance Sampling"

$$w_i = \frac{p(\theta^{(i)})}{p(\theta^{(i)}|D)} = \frac{\cancel{p(\theta^{(i)})}\, p(D)}{p(D|\theta^{(i)})\, \cancel{p(\theta^{(i)})}}$$

# Monte Carlo Estimates of $p(D|M)$ (cont.)

$$\hat{p}(D) = \frac{\frac{1}{m}\sum_{i=1}^{m} \frac{p(D)}{p(D|\theta^{(i)})} p(D|\theta^{(i)})}{\sum_{i=1}^{m} \frac{p(D)}{p(D|\theta^{(i)})}}$$

$$= \left\{ \frac{1}{m} \sum_{i=1}^{m} p(D|\theta^{(i)})^{-1} \right\}^{-1}$$

Newton and Raftery's "Harmonic Mean Estimator"

Unstable in practice and needs modification

# $p(D|M)$ from Gibbs Sampler Output

First note the following identity (for any $\theta^*$):

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

$p(D|\theta^*)$ and $p(\theta^*)$ are usually easy to evaluate.

What about $p(\theta^*|D)$?

Suppose we decompose $\theta$ into $(\theta_1,\theta_2)$ such that $p(\theta_1|D,\theta_2)$ and $p(\theta_2|D,\theta_1)$ are available in closed-form…

Chib (1995)

# $p(D|M)$ from Gibbs Sampler Output

$$p(\theta_1^*, \theta_2^* \mid D) = p(\theta_2^* \mid D, \theta_1^*)\, p(\theta_1^* \mid D)$$

The Gibbs sampler gives (dependent) draws from $p(\theta_1, \theta_2 \mid D)$ and hence marginally from $p(\theta_2 \mid D)$…

$$p(\theta_1^* \mid D) = \int p(\theta_1^* \mid D, \theta_2)\, p(\theta_2 \mid D)\, d\theta_2$$

$$\approx \frac{1}{G} \sum_{g=1}^{G} p(\theta_1^* \mid D, \theta_2^{(g)})$$

# What about 3 parameter blocks…

$$p(\theta_1^*, \theta_2^*, \theta_3^* \mid D) = p(\theta_3^* \mid D, \theta_1^*, \theta_2^*)\, p(\theta_2^* \mid D, \theta_1^*)\, p(\theta_1^* \mid D)$$

OK ? OK

$$p(\theta_2^* \mid D, \theta_1^*) = \int p(\theta_2^* \mid D, \theta_1^*, \theta_3)\, p(\theta_3 \mid D, \theta_1^*)\, d\theta_3$$

$$\approx \frac{1}{G} \sum_{g=1}^{G} p(\theta_2^* \mid D, \theta_1^*, \theta_3^{(g)})$$

To get these draws, continue the Gibbs sampler sampling in turn from:

$$p(\theta_2 \mid D, \theta_1^*, \theta_3) \quad \text{and} \quad p(\theta_3 \mid D, \theta_1^*, \theta_2)$$

# Bayesian Information Criterion

$$S_{BIC}(M_k) = -2\text{loglik} + p_k \log n, \quad k = 1,\ldots,K$$

- BIC is an O(1) "approximation" to $p(D|M)$
- Circumvents explicit prior
- Approximation is $O(n^{-1/2})$ for a class of priors called "unit information priors."
- No free lunch (Weakliem (1998) example)