COMS 4721: Machine Learning for Data Science

Lecture 24, 4/25/2017

Prof. John Paisley

Department of Electrical Engineering
& Data Science Institute

Columbia University

Model Selection

The model selection problem

We've seen how often model parameters need to be set in advance and discussed how this can be done using using cross-validation.

Another type of model selection problem is learning model order.

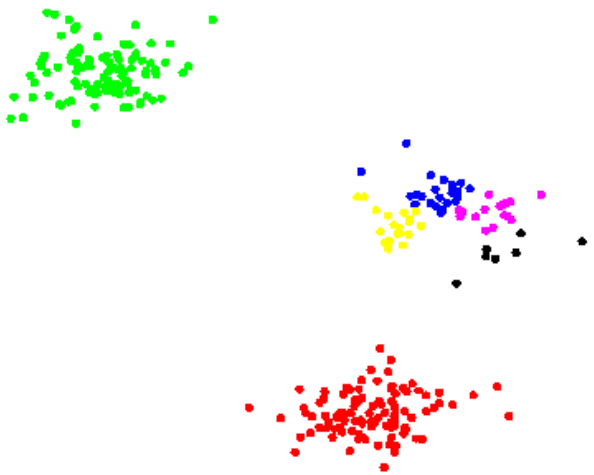
Model order: The complexity of a class of models

- Gaussian mixture model: How many Gaussians?
- Matrix factorization: What rank?
- Hidden Markov models: How many states?

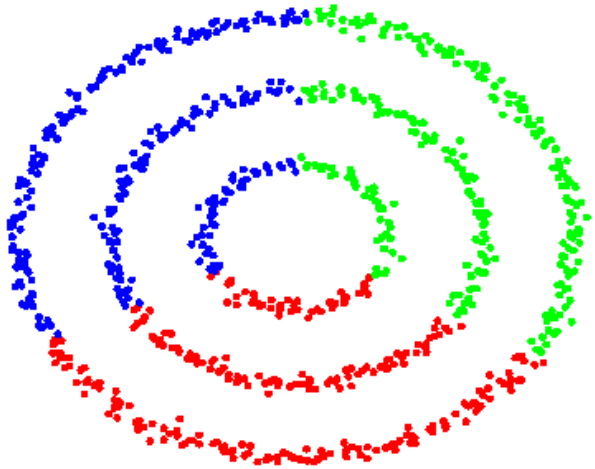
In each of these problems, we can't simply look at the log-likelihood because a more complex model can always fit the data better.

Model Order

We will discuss two methods for selecting an "appropriate" complexity of the model. This assumes a good model type was chosen to begin with.



(a) Inappropriate model order.



(b) Inappropriate model type.

Example: Maximum likelihood

Notation

We write $\mathcal{L}$ for the log-likelihood of a parameter under a model $p(x|\theta)$:

$$x_i \overset{iid}{\sim} p(x|\theta) \quad \Longleftrightarrow \quad \mathcal{L} = \sum_{i=1}^{N} \log p(x_i|\theta)$$

The maximum likelihood solution is: $\theta_{\text{ML}} = \arg\max_\theta \mathcal{L}$.

Example: How many clusters? (wrong way)

The parameters θ could be those of a GMM. We could find θ_{ML} for different numbers of clusters and pick the one with the largest $\mathcal{L}$.

Problem: We can perfectly fit the data by putting each observation in its own cluster. Then shrink the variance of each Gaussian to zero.

Number of Parameters

The general problem

- Models with more degrees of freedom are more prone to overfitting.
- The degrees of freedom is roughly the number of scalar parameters, K.
- By increasing K (done by increasing #clusters, rank, #states, etc.) the model can add more degrees of freedom.

Some common solutions

- **Stability**: Bootstrap sample the data, learn a model, calculate the likelihood on the original data set. Repeat and pick the best model.
- **Bayesian nonparametric methods**: Each possible value of K is assigned a prior probability. The posterior learns the best K.
- **Penalization approaches**: A penalty term makes adding parameters expensive. Must be overcome by a greater improvement in likelihood.

General form

Define a *penalty function* on the number of model parameters. Instead of maximizing $\mathcal{L}$, minimize $-\mathcal{L}$ and add the defined penalty.

Two popular penalties are:

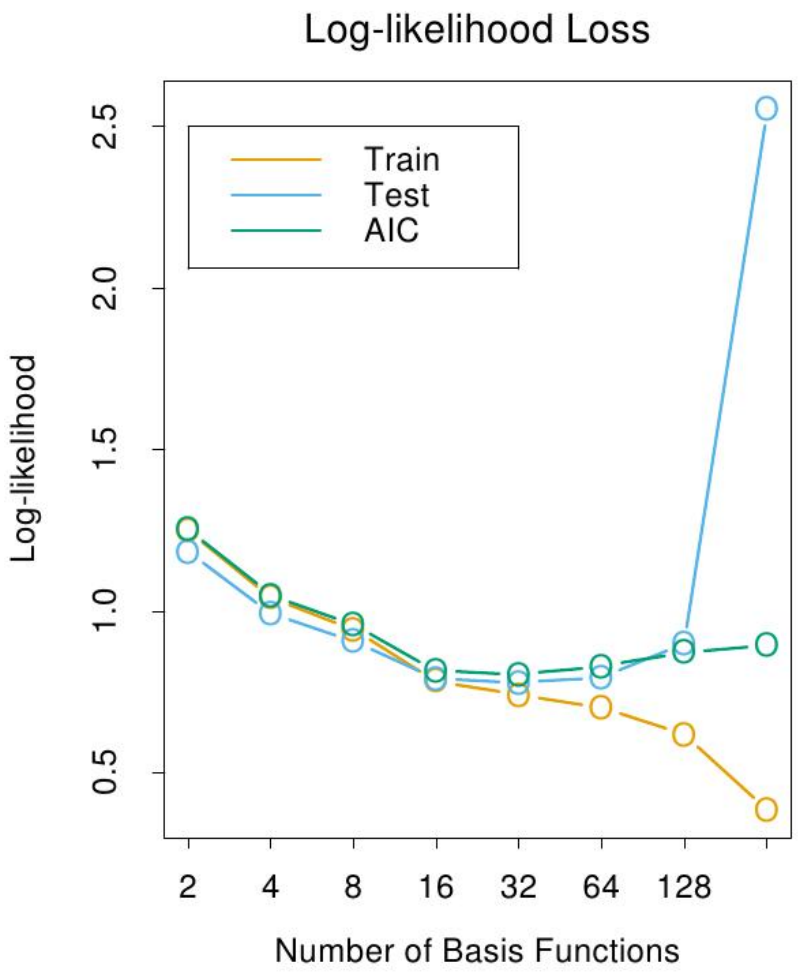
- **Akaike information criterion (AIC)**: $-\mathcal{L} + K$
- **Bayesian information criterion (BIC)**: $-\mathcal{L} + \frac{1}{2} K \ln N$

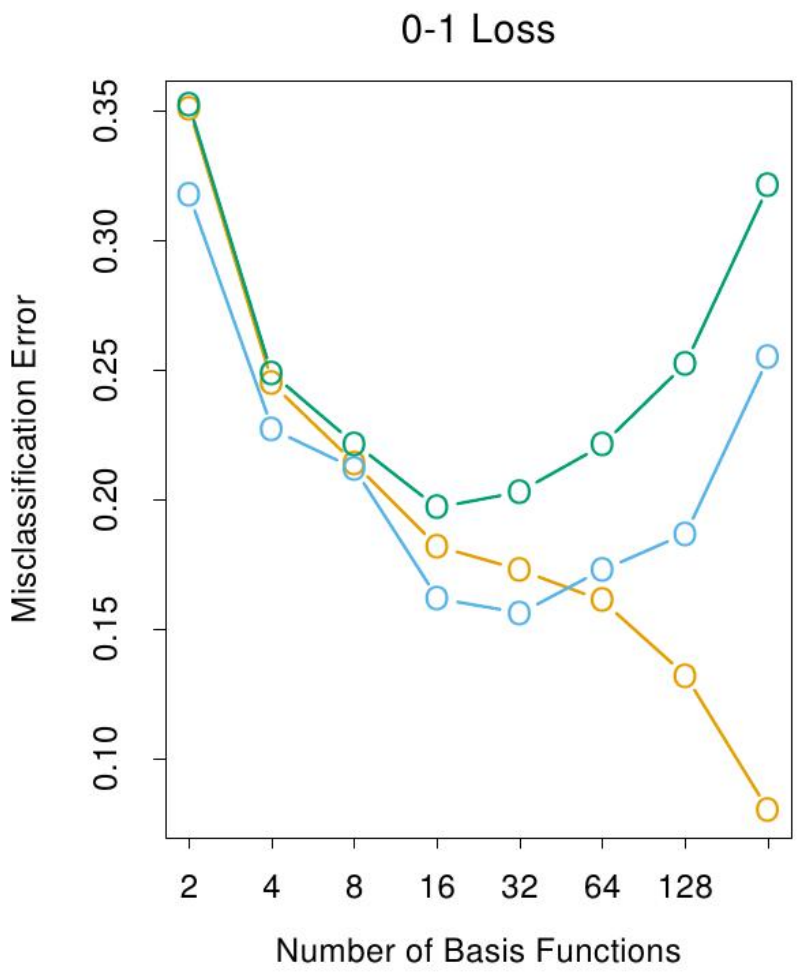
When $\frac{1}{2} \ln N > 1$, BIC encourages a simpler model (happens when $N \geq 8$).

Example: For NMF with an $M_1 \times M_2$ matrix and rank R factorization,

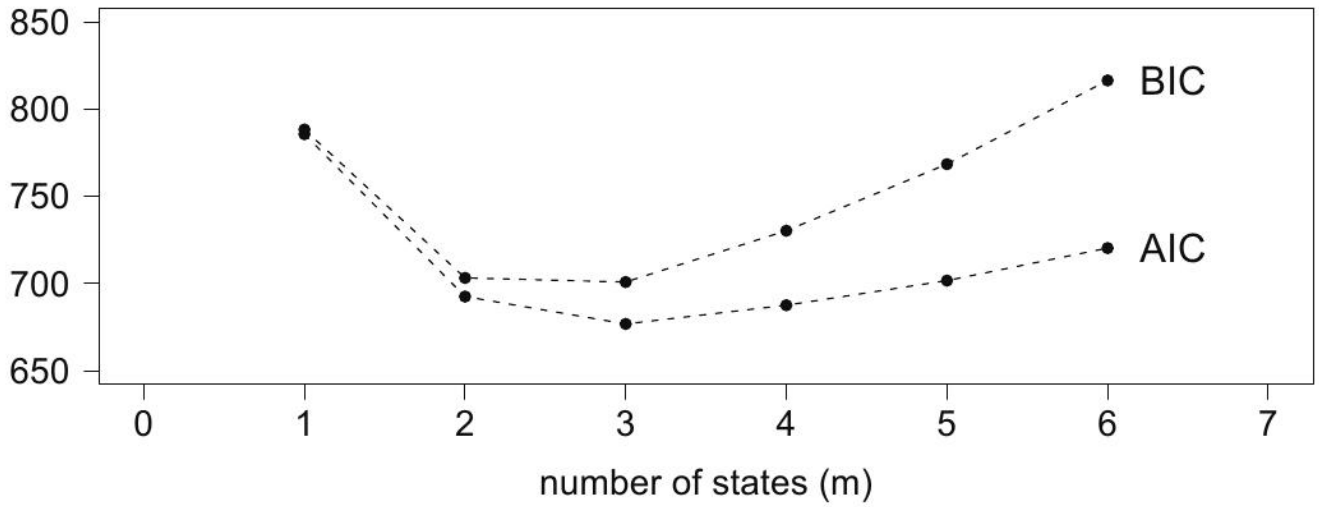
$$\text{AIC} \;\to\; (M_1 + M_2)R, \qquad \text{BIC} \;\to\; \frac{1}{2}(M_1 + M_2)R \ln(M_1 M_2)$$

Example of AIC output





EXAMPLE: AIC VS BIC ON HMM



model	$-\log L$	AIC	BIC
'1-state HM'	391.9189	785.8	788.5
2-state HM	342.3183	692.6	703.3
3-state HM	329.4603	**676.9**	**701.0**
4-state HM	327.8316	687.7	730.4
5-state HM	325.9000	701.8	768.6
6-state HM	324.2270	720.5	816.7
indep. mixture (2)	360.3690	726.7	734.8
indep. mixture (3)	356.8489	723.7	737.1
indep. mixture (4)	356.7337	727.5	746.2

Notice:

- Likelihood is always improving
- Only compare location of AIC and BIC minima, not the values.

Derivation of BIC

AIC AND BIC

Recall the two penalties:

- **Akaike information criterion (AIC)**: $-\mathcal{L} + K$
- **Bayesian information criterion (BIC)**: $-\mathcal{L} + \frac{1}{2} K \ln N$

Algorithmically, there is no extra work required:

1. Find the ML solution of the selected models and calculate $\mathcal{L}$.
2. Add the AIC or BIC penalty to get a score useful for picking a model.

Q: Where do these penalties come from? Currently they seem arbitrary.

A: We will derive BIC next. AIC also has a theoretical motivation, but we will not discuss that derivation.

DERIVING THE BIC

Imagine we have r candidate models, $\mathcal{M}_1, \ldots, \mathcal{M}_r$. For example, r HMMs each having a different number of states.

We also have data $\mathcal{D} = \{x_1, \ldots, x_N\}$. We want the posterior of each $\mathcal{M}_i$.

$$p(\mathcal{M}_i|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M}_i)p(\mathcal{M}_i)}{\sum_j p(\mathcal{D}|\mathcal{M}_j)p(\mathcal{M}_j)}$$

If we assume a uniform prior distribution on models, then because the denominator is constant in $\mathcal{M}_i$, we can pick

$$\mathcal{M} = \arg\max_{\mathcal{M}_i} \ln p(\mathcal{D}|\mathcal{M}_i) = \int \ln p(\mathcal{D}|\theta, \mathcal{M}_i)p(\theta|\mathcal{M}_i)d\theta$$

We're choosing the model with the largest *marginal likelihood* of the data by integrating out all parameters of the model. This is usually not solvable.

DERIVING THE BIC

We will see how the BIC arises from the approximation,

$$\mathcal{M} = \arg\max_{\mathcal{M}_i} \ln p(\mathcal{D}|\mathcal{M}_i) \approx \arg\max_{\mathcal{M}_i} \ln p(\mathcal{D}|\theta_{\mathrm{ML}}, \mathcal{M}_i) - \frac{1}{2} K \ln N$$

Step 1: Recognize that the difficulty is with the integral

$$\ln p(\mathcal{D}|\mathcal{M}_i) = \ln \int p(\mathcal{D}|\theta) p(\theta) d\theta.$$

$\mathcal{M}_i$ determines $p(\mathcal{D}|\theta)$, $p(\theta)$—we will suppress this conditioning.

Step 2: Approximate this integral using a second-order Taylor expansion.

DERIVING THE BIC

1. We want to calculate:

$$\ln p(\mathcal{D}|\mathcal{M}) \;=\; \ln \int p(\mathcal{D}|\theta)p(\theta)d\theta \;=\; \ln \int \exp\{\ln p(\mathcal{D}|\theta)\}p(\theta)d\theta$$

2. We use a second-order Taylor expansion of $\ln p(\mathcal{D}|\theta)$ at the point θ_{ML},

$$\begin{aligned}\ln p(\mathcal{D}|\theta) \quad \approx \quad & \ln p(\mathcal{D}|\theta_{\text{ML}}) + (\theta - \theta_{\text{ML}})^T \underbrace{\nabla \ln p(\mathcal{D}|\theta_{\text{ML}})}_{=0} \\ & + \frac{1}{2}(\theta - \theta_{\text{ML}})^T \underbrace{\nabla^2 \ln p(\mathcal{D}|\theta_{\text{ML}})}_{=-\mathcal{J}(\theta_{\text{ML}})}(\theta - \theta_{\text{ML}})\end{aligned}$$

3. Approximate $p(\theta)$ as uniform and plug this approximation back in,

$$\ln p(\mathcal{D}|\mathcal{M}) \;\approx\; \ln p(\mathcal{D}|\theta_{\text{ML}}) + \ln \int \exp\left\{-\frac{1}{2}(\theta-\theta_{\text{ML}})^T \mathcal{J}(\theta_{\text{ML}})(\theta-\theta_{\text{ML}})\right\}d\theta$$

DERIVING THE BIC

Observation: The integral is the normalizing constant of a Gaussian,

$$\int \exp\Big\{ -\frac{1}{2}(\theta-\theta_{\mathrm{ML}})^T \mathcal{J}(\theta_{\mathrm{ML}})(\theta-\theta_{\mathrm{ML}})\Big\} d\theta \;=\; \left(\frac{2\pi}{|\mathcal{J}(\theta_{\mathrm{ML}})|}\right)^{K/2}$$

Remember the definition that

$$-\mathcal{J}(\theta_{\mathrm{ML}}) \;=\; \nabla^2 \ln p(\mathcal{D}|\theta_{\mathrm{ML}}) \;\overset{(a)}{=}\; N \underbrace{\sum_{i=1}^{N} \frac{1}{N} \nabla^2 \ln p(x_i|\theta_{\mathrm{ML}})}_{\text{converges as } N \text{ increases}}$$

(a) is by the i.i.d. model assumption made at the beginning of the lecture.

Deriving the BIC

4. Plugging this in,

$$\ln p(\mathcal{D}|\mathcal{M}) \;\approx\; \ln p(\mathcal{D}|\theta_{\mathrm{ML}}) \;+\; \ln\left(\frac{2\pi}{|\mathcal{J}(\theta_{\mathrm{ML}})|}\right)^{K/2}$$

and $|\mathcal{J}(\theta_{\mathrm{ML}})| = N\left|\sum_{i=1}^{N}\frac{1}{N}\nabla^2 \ln p(x_i|\theta_{\mathrm{ML}})\right|$.

Therefore we arrive at the BIC,

$$\ln p(\mathcal{D}|\mathcal{M}) \;\approx\; \ln p(\mathcal{D}|\theta_{\mathrm{ML}}) \;-\; \frac{1}{2}K\ln N \;+\; \underbrace{\text{something not growing with } N}_{O(1)\text{ term, so we ignore it}}$$

Some next steps

ICML Sessions (subset)

The International Conference on Machine Learning (ICML) is a major ML conference. Many of the session titles should look familiar:

- Bayesian Optimization and Gaussian Processes
- PCA and Subspace Models
- Supervised Learning
- Matrix Completion and Graphs
- Clustering and Nonparametrics
- Active Learning
- Clustering
- Boosting and Ensemble Methods
- Matrix Factorization I & II
- Kernel Methods I & II
- Topic models
- Time Series and Sequences
- etc.

ICML Sessions (subset)

Other sessions might not look so familiar:

- Reinforcement Learning I & II
- Bandits I & II
- Optimization I, II & III
- Bayesian nonparametrics I & II
- Online learning I & II
- Graphical Models I & II
- Neural Networks and Deep Learning I & II
- Metric Learning and Feature Selection
- etc.

Many of these topics are taught in advanced machine learning courses at Columbia in the CS, Statistics, IEOR and EE departments.