Lecture 20: Model Order Selection, Exponential Family Models

GU4241/GR5241 Statistical Machine Learning

Linxi Liu

April 13, 2018

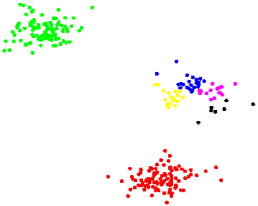
Model Selection for Clustering

The model selection problem

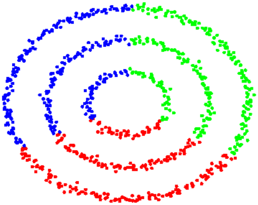
For mixture models $\pi(x) = \sum_{k=1}^{K} c_k p(x|\theta_k)$, we have so far assumed that the number K of clusters is known.

Model Order

Methods which automatically determine the complexity of a model are called **model selection** methods. The number of clusters in a mixture model is also called the **order** of the mixture model, and determining it is called **model order selection**.



(a) Inappropriate model order.



(b) Inappropriate model type.

Model Selection for Clustering

Notation

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

$$\mathcal{L}(\mathbf{x}^n; \theta) := \log \prod_{i=1}^{n} p(x_i|\theta)$$

In particular, for a mixture model:

$$\mathcal{L}(\mathbf{x}^n; \mathbf{c}, \boldsymbol{\theta}) := \log \prod_{i=1}^{n} \Big(\sum_{k=1}^{K} c_k p(x_i|\theta_k) \Big)$$

Number of clusters: Naive solution (wrong!)

We could treat K as a parameter and use maximum likelihood, i.e. try to solve:

$$(K, c_1, \ldots, c_K, \theta_1, \ldots, \theta_K) := \operatorname{arg} \max_{K, \mathbf{c}', \boldsymbol{\theta}'} \mathcal{L}(\mathbf{x}^n; K, \mathbf{c}', \boldsymbol{\theta}')$$

Number of Clusters

Problem with naive solution: Example

Suppose we use a Gaussian mixture model.

- The optimization procedure can add additional components arbitrarily.
- It can achieve minimal fitting error by using a separate mixture component for each data point (ie $\mu_k = x_i$).
- By reducing the variance of each component, it can additionally increase the density value at $\mu_k = x_i$. That means we can achieve arbitrarily high log-likelihood.
- Note that such a model (with very high, narrow component densities at the data points) would achieve *low* log-likelihood on a new sample from the same source. In other words, it does not generalize well.

In short: The model overfits.

Number of Clusters

The general problem

- Recall our discussion of model complexity: Models with more degrees of freedom are more prone to overfitting.
- The number of degrees of freedom is roughly the number of scalar parameters.
- By increasing K, the clustering model can *add more degrees of freedom*.

Most common solutions

- **Penalization approaches**: A penalty term makes adding parameters expensive. Similar to shrinkage in regression.
- **Stability**: Perturb the distribution using resampling or subsampling. Idea: A choice of K for which solutions are stable under perturbation is a good explanation of the data.
- **Bayesian methods**: Each possible value of K is assigned a probability, which is combined with the likelihood given K to evaluate the plausibility of the solution. Somewhat related to penalization.

Penalization Strategies

General form

Penalization approaches define a *penalty function* ϕ, which is an increasing function of the number m of model parameters.
Instead of *maximizing* the log-likelihood, we *minimize* the *negative* log-likelihood and add ϕ:

$$(m, \theta_1, \ldots, \theta_m) = \mathsf{arg} \min_{m, \theta_1, \ldots, \theta_m} -\mathcal{L}(\mathbf{x}^n; \theta_1, \ldots, \theta_m) + \phi(m)$$

The most popular choices

The penalty function

$$\phi_{\mathsf{AIC}}(m) := m$$

is the **Akaike information criterion (AIC)**.

$$\phi_{\mathsf{BIC}}(m) := \frac{1}{2} m \log n$$

is the **Bayesian information criterion (BIC)**.

Clustering

Clustering with penalization

For clustering, AIC means:

$$(K, \mathbf{c}, \boldsymbol{\theta}) = \text{arg} \min_{K,\mathbf{c}',\boldsymbol{\theta}'} -\mathcal{L}(\mathbf{x}^n; K, \mathbf{c}', \boldsymbol{\theta}') + K$$

Similarly, BIC solves:

$$(K, \mathbf{c}, \boldsymbol{\theta}) = \text{arg} \min_{K,\mathbf{c}',\boldsymbol{\theta}'} -\mathcal{L}(\mathbf{x}^n; K, \mathbf{c}', \boldsymbol{\theta}') + \frac{1}{2} K \log n$$

Which criterion should we use?

- BIC penalizes additional parameters more heavily than AIC (i.e. tends to select fewer components).
- Various theoretical results provide conditions under which one of the criteria succeeds or fails, depending on:
  - Whether the sample is small or large.
  - Whether the individual components are mispecified or not.
- BIC is more common choice in practice.

Stability

Assumption

A value of K is plausible if it results in similar solutions on separate samples.

Strategy

As in cross validation and boostrap methods, we "simulate" different sample sets by perturbation or random splits of the input data.

Recall: Assignment in mixtures

Recall that, under a mixture model $\pi = \sum_{k=1}^{K} c_k p(x|\theta_k)$, we compute a "hard" assignment for a data point x_i as

$$m_i := \arg\max_k c_k p(x_i|\theta_k)$$

Stability

Computing the stability score for fixed K

1. Randomly split the data into two sets $\mathcal{X}'$ and $\mathcal{X}''$ of equal size.
2. Separately estimate mixture models π' on $\mathcal{X}'$ and π'' on $\mathcal{X}''$, using EM.
3. For each data point $x_i \in \mathcal{X}''$, compute assignments m_i' under π' and m_i'' under π''. (That is: π' is now used for prediction on $\mathcal{X}''$.)
4. Compute the score

$$\psi(K) := \min_{\sigma} \sum_{i=1}^{n} \mathbb{I}\{m_i' \neq \sigma(m_i'')\}$$

where the minimum is over all permutations σ which permute $\{1, \ldots, K\}$.

Stability

Explanation

- $\psi(K)$ measures: How many points are assigned to a different cluster under π' than under π''?
- The minimum over permutations is necessary because the numbering of clusters is not unique. (Cluster 1 in π' might correspond to cluster 5 in π'', etc.)

Stability

Selecting the number of clusters

1. Compute $\psi(K)$ for a range of values of K.
2. Select K for which $\psi(K)$ is minimial.

Improving the estimate of $\psi(K)$

For each K, we can perform multiple random splits and estimate $\psi(K)$ by averaging over these.

Performance

- Empirical studies show good results on a range of problems.
- Some basic theoretical results available, but not as detailed as for AIC or BIC.

Exponential Family Distributions

Definition

We consider a model $\mathcal{P}$ for data in a sample space $\mathbf{X}$ with parameter space $\mathcal{T} \subset \mathbb{R}^m$. Each distribution in $\mathcal{P}$ has density $p(x|\theta)$ for some $\theta \in \mathcal{T}$.

The model is called an **exponential family model** (EFM) if p can be written as

$$p(x|\theta) = \frac{h(x)}{Z(\theta)} e^{\langle S(x), \theta \rangle}$$

where:

- S is a function $S : \mathbf{X} \to \mathbb{R}^m$. This function is called the **sufficient statistic** of $\mathcal{P}$.
- h is a function $h : \mathbf{X} \to \mathbb{R}_+$.
- Z is a function $Z : \mathcal{T} \to \mathbb{R}_+$, called the **partition function**.

Exponential Family Distributions

Exponential families are important because:

1. The special form of p gives them many nice properties.
2. Most important parametric models (e.g. Gaussians) are EFMs.
3. Many algorithms and methods can be formulated generically for all EFMs.

Alternative Form

The choice of p looks perhaps less arbitrary if we write

$$p(x|\theta) = \exp\Big(\langle S(x), \theta\rangle - \phi(x) - \psi(\theta)\Big)$$

which is obtained by defining

$$\phi(x) := -\log(h(x)) \qquad \text{and} \qquad \psi(\theta) := \log(Z(\theta))$$

A first interpretation

Exponential family models are models in which:

- The data and the parameter interact only through the linear term $\langle S(x), \theta\rangle$ in the exponent.

Alternative Form

The choice of p looks perhaps less arbitrary if we write

$$p(x|\theta) = \exp\Big(\langle S(x), \theta\rangle - \phi(x) - \psi(\theta)\Big)$$

which is obtained by defining

$$\phi(x) := -\log(h(x)) \qquad \text{and} \qquad \psi(\theta) := \log(Z(\theta))$$

A first interpretation

Exponential family models are models in which:

- The data and the parameter interact only through the linear term $\langle S(x), \theta\rangle$ in the exponent.
- The logarithm of p can be non-linear in both $S(x)$ and θ, but there is no *joint* nonlinear function of $(S(x), \theta)$.

The Partition Function

Normalization constraint

Since p is a probability density, we know

$$\int_{\mathbf{X}} \frac{h(x)}{Z(\theta)} e^{\langle S(x),\theta \rangle} dx = 1 \; .$$

Partition function

The only term we can pull out of the integral is the partition function $Z(\theta)$, hence

$$Z(\theta) = \int_{\mathbf{X}} h(x) e^{\langle S(x),\theta \rangle} dx$$

Note: This implies that an exponential family is completely determined by choice of the spaces $\mathbf{X}$ and $\mathcal{T}$ and of the functions S and h.

Example: Gaussian

In 1 dimension

We can rewrite the exponent of the Gaussian as

$$\begin{aligned}\frac{1}{\sqrt{2\pi}\sigma}\exp\Big(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\Big) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\Big(-\frac{1}{2}\frac{x^2}{\sigma^2}+\frac{2x\mu}{2\sigma^2}\Big)\exp\Big(-\frac{1}{2}\frac{\mu^2}{\sigma^2}\Big)\\ &= \underbrace{c(\mu,\sigma)}_{\text{some function of } \mu \text{ and } \sigma}\exp\Big(x^2\cdot\frac{-1}{2\sigma^2}+x\cdot\frac{\mu}{\sigma^2}\Big)\end{aligned}$$

This shows the Gaussian is an exponential family, since we can choose:

$S(x) := \big(x^2, x\big)$ and $\theta := \big(\frac{-1}{2\sigma^2}, \frac{\mu}{\sigma^2}\big)$ and $h(x) = 1$ and $Z(\theta) = c(\mu,\sigma)^{-1}$.

In d dimensions

$$S(\mathbf{x}) = \big(\mathbf{x}\mathbf{x}^t, \mathbf{x}\big) \qquad \text{and} \qquad \theta := \big(-\tfrac{1}{2}\Sigma^{-1}, \Sigma^{-1}\mu\big)$$

More Examples of Exponential Families

Model	Sample space	Sufficient statistic
Gaussian	$\mathbb{R}^d$	$S(\mathbf{x}) = (\mathbf{x}\mathbf{x}^t, \mathbf{x})$
Gamma	$\mathbb{R}_+$	$S(x) = (\ln(x), x)$
Poisson	$\mathbb{N}_0$	$S(x) = x$
Multinomial	$\{1, \dots, K\}$	$S(x) = x$
Wishart	Positive definite matrices	(requires more details)
Mallows	Rankings (permutations)	(requires more details)
Beta	$[0, 1]$	$S(x) = (\ln(x), \ln(1 - x))$
Dirichlet	Probability distributions on d events	$S(\mathbf{x}) = (\ln x_1, \dots, \ln x_d)$
Bernoulli	$\{0, 1\}$	$S(x) = x$
...	...	...

Roughly speaking

On every sample space, there is a "natural" statistic of interest. On a space with Euclidean distance, for example, it is natural to measure both location *and* correlation; on categories (which have no "distance" from each other), it is more natural to measure only expected numbers of counts.
On most types of sample spaces, the exponential family model with S chosen as this natural statistic is the prototypical distribution.

Maximum Likelihood for EFMs

Log-likelihood for n samples

$$\log \prod_{i=1}^{n} p(x_i|\theta) = \sum_{i=1}^{n} \Big(\log(h(x_i)) - \log(Z(\theta)) + \langle S(x_i), \theta \rangle \Big)$$

MLE equation

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta} \sum_{i=1}^{n} \Big(\log(h(x_i)) - \log(Z(\theta)) + \langle S(x_i), \theta \rangle \Big) \\ &= -n \frac{\partial}{\partial \theta} \log(Z(\theta)) + \sum_{i=1}^{n} S(x_i) \end{aligned}$$

Hence, the MLE is the parameter value $\hat{\theta}$ which satisfies the equation

$$\frac{\partial}{\partial \theta} \log(Z(\hat{\theta})) = \frac{1}{n} \sum_{i=1}^{n} S(x_i)$$

Moment Matching

Further simplification

We know that $Z(\theta) = \int h(x) \exp \langle S(x), \theta \rangle \, dx$, so

$$\frac{\partial}{\partial \theta} \log(Z(\theta)) = \frac{\frac{\partial}{\partial \theta} Z(\theta)}{Z(\theta)} = \frac{\int h(x) \frac{\partial}{\partial \theta} e^{\langle S(x), \theta \rangle} dx}{Z(\theta)} = \frac{\int S(x) h(x) e^{\langle S(x), \theta \rangle} dx}{Z(\theta)} = \mathbb{E}_{p(x|\theta)}[S(x)]$$

MLE equation

Substitution into the MLE equation shows that $\hat{\theta}$ is given by

$$\mathbb{E}_{p(x|\hat{\theta})}[S(x)] = \frac{1}{n} \sum_{i=1}^{n} S(x_i)$$

Using the empirical distribution $\mathbb{F}_n$, the right-hand side can be expressed as

$$\mathbb{E}_{p(x|\hat{\theta})}[S(x)] = \mathbb{E}_{\mathbb{F}_n}[S(x)]$$

This is called a **moment matching equation**. Hence, MLEs of exponential family models can be obtained by moment matching.

Summary: MLE for EFMs

The MLE

If $p(x|\theta)$ is an exponential family model with sufficient statistic S, the maximum likelihood estimator $\hat{\theta}$ of θ given data $x_1, \ldots, x_n$ is given by the equation

$$\mathbb{E}_{p(x|\hat{\theta})}[S(x)] = \frac{1}{n} \sum_{i=1}^{n} S(x_i)$$

Note

We had already noticed that the MLE (for some parameter τ) is often of the form

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} f(x_i) \ .$$

Models are often defined so that the parameters can be interpreted as expectations of some useful statistic (e.g., a mean or variance). If θ in an exponential family is chosen as $\theta = \mathbb{E}_{p(x|\theta)}[S(x)]$, then we have indeed

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} S(x_i) \ .$$

EM for Exponential Family Mixture

Finite mixture model

$$\pi(x) = \sum_{k=1}^{K} c_k p(x|\theta_k) \ ,$$

where p is an exponential family with sufficient statistic S.

EM Algorithm

- **E-Step:** Recompute the assignment weight matrix as

$$a_{ik}^{(\mathsf{j+1})} := \frac{c_k^{(\mathsf{j})} p(x_i|\theta_k^{(\mathsf{j})})}{\sum_{l=1}^{K} c_l^{(\mathsf{j})} p(x_i|\theta_l^{(\mathsf{j})})} \ .$$

- **M-Step:** Recompute the proportions c_k and parameters θ_k by solving

$$c_k^{(\mathsf{j+1})} := \frac{\sum_{i=1}^{n} a_{ik}^{(\mathsf{j+1})}}{n} \qquad \text{and} \qquad \mathbb{E}_{p(x|\theta_k^{(\mathsf{j+1})})}[S(x)] = \frac{\sum_{i=1}^{n} a_{ik}^{(\mathsf{j+1})} S(x_i)}{\sum_{i=1}^{n} a_{ik}^{(\mathsf{j+1})}}$$

EM for Exponential Family Mixture

If in particular the model is parameterized such that

$$\mathbb{E}_{p(x|\theta)}[S(x)] = \theta$$

the algorithm becomes very simple:

- **E-Step:** Recompute the assignment weight matrix as

$$a_{ik}^{(\mathsf{j+1})} := \frac{c_k^{(\mathsf{j})} p(x_i|\theta_k^{(\mathsf{j})})}{\sum_{l=1}^{K} c_l^{(\mathsf{j})} p(x_i|\theta_l^{(\mathsf{j})})} \ .$$

- **M-Step:** Recompute the proportions c_k and parameters θ_k as

$$c_k^{(\mathsf{j+1})} := \frac{\sum_{i=1}^{n} a_{ik}^{(\mathsf{j+1})}}{n} \qquad \text{and} \qquad \theta_k^{(\mathsf{j+1})} := \frac{\sum_{i=1}^{n} a_{ik}^{(\mathsf{j+1})} S(x_i)}{\sum_{i=1}^{n} a_{ik}^{(\mathsf{j+1})}}$$