Finite Mixture Models

Jasmine. Hao[1]

[1]University of Hong Kong

ECON 6083: Machine Learning

Outline

Machine Learning

- Supervised:
  - We are given input/output samples (X, y) which we relate with a function $y = f(X)$.
  - We would like to "learn" f, and evaluate it on new data.
  - Types:
    - Classification: y is discrete (class labels).
    - Regression: y is continuous, e.g., linear regression.
- Unsupervised:
  - Given only samples X of the data, we compute a function f such that $y = f(X)$ is "simpler".
  - Types:
    - Clustering: y is discrete.
    - y is continuous: Matrix factorization, Kalman filtering, unsupervised neural networks.

Machine Learning

- **Unsupervised:**
  - Cluster some hand-written digit data into 10 classes.
  - What are the top 20 topics in Twitter right now?
  - Find and cluster distinct accents of people at Berkeley.

Methods

- **Supervised Learning:**
  - kNN (k-th Nearest Neighbors)
  - Naïve Bayes
  - Linear/Logistic Regression
  - Support Vector Machines
  - Random Forests
  - Neural Networks
- **Unsupervised Learning:**
  - Clustering(k-Means)
  - Mixture Models

A General Model for Mixture Distribution

- **Key Equation:**

$$F(y) = \int F(y|\alpha) dG(\alpha) \tag{1}$$

- **Definitions:**
  - **Observed Random Variable** $y \in Y$: The variable that we observe.
  - **Latent Variable** $\alpha \in A$: An unobserved or mixing variable; A could be infinite.
  - **Mixture Distribution** F: The cumulative distribution function (cdf) of Y.
  - **Component Distribution** $F(\cdot|\alpha)$: A collection of component cdfs indexed by α.
  - **Mixing Distribution** G: A cdf over the space A, determining how components combine.
  - When α takes a finite number of values, the weight G attaches to each α_i is known as the **weight of component** i.
- **Objective:** Estimate the unknown parameters $F(\cdot|\alpha)$ and G using repeated observations $(y_i)_{i=1}^n$ from F.

Pros and Cons for Mixture Models

- Pros: **increase the flexibility of the models** and **reduce their reliance on restrictive parametric assumptions**.
  - In parametric approaches, model misspecification often leads to inconsistent estimates and invalid inference.
  - Proportional hazard models (Heckman & Singer (1984)); discrete choice models (Matzkin (1992)); switching regression models (Kitamura (2003)).
- Cons: **reduce identifying power**.
  - Relaxing parametric assumptions may not allow unique identification of the parameters of interest, resulting in **partial identification**.
- We consider identification issues for taking semi-& non-parametric approaches to mixture models.

Mixtures of Gaussians and the EM Algorithm

- **Modeling Data**:
  - Joint distribution for y_i(scalar): $p(y_i, \alpha_i) = p(y_i|\alpha_i)p(\alpha_i)$.
  - Latent variables $\alpha_i \sim$ Multinomial(ϕ).
  - Conditionals $y_i|\alpha_i = j \sim \mathcal{N}(\mu_j, \sigma_j)$.
- **Mixture of Gaussians Model**:
  - Represents data as a mixture from k different Gaussian distributions.
  - ϕ_j controls the probability that α_i takes on value j.
  - The α_i's are latent variables, making estimation challenging.
- **Latent Variables**:
  - α_i's are not directly observed (*hidden*).
  - They represent the unobserved "origin" of each data point y_i.
- **EM Algorithm for Density Estimation**:
  - Designed to model data without labels (*unsupervised learning*).

Outline

1. Supervised v.s. Unsupervised Learning
2. Overview of Mixture Models
3. Mixture Gaussian and EM algorithm
   - Identification Challenge
4. Finite Mixtures Variations

MLE for Normal Variable I

- The probability density function of a normal distribution for a single observation y_i is given by:

$$f(y_i|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right)$$

- The likelihood function L for the entire sample is the product of the individual densities:

$$L(\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right)$$

- The log-likelihood function simplifies the multiplication of probabilities into a sum:

$$l(\mu,\sigma^2) = \log L(\mu,\sigma^2) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right)\right)$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2$$

MLE for Normal Variable II

- To find the MLEs, we take the derivatives with respect to μ and σ^2 and set them to zero:
  - **Derivative with respect to μ:**

$$\frac{\partial}{\partial \mu} \log L(\mu, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \mu) = 0$$

Solving for μ, we find:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

This is the sample mean.

MLE for Normal Variable III

- **Derivative with respect to σ^2:**

$$\frac{\partial}{\partial \sigma^2} \log L(\mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (y_i - \mu)^2 = 0$$

Solving for σ^2, we find:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\mu})^2$$

This is the sample variance.

- The maximum likelihood estimators for the mean μ and variance σ^2 of a normally distributed sample are $\hat{\mu}$, the sample mean, and $\hat{\sigma}^2$, the sample variance, respectively. These estimators are widely used due to their desirable statistical properties.

Parameter Estimation

- Now, consider a 2-type mixture model, with ϕ being $p(\alpha_i = 1)$. The parameters of our model are μ_1, μ_2, σ_1, σ_2 and ϕ.
- **Likelihood of Data** Let the j denote the type and i denote the sample index. To estimate these parameters, the likelihood of our data is expressed as:

$$\begin{aligned}
\ell(\phi, \mu, \Sigma) &= \log\left(\sum_{j=1}^{2} f(y_i|\mu_j, \sigma_j^2) p(\alpha_i = j; \phi)\right) \\
&= \log\left(\phi f(y_i|\mu_1, \sigma_1^2) + (1-\phi) f(y_i|\mu_2, \sigma_2^2)\right) \\
&= \log\left(\phi \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(y_i-\mu_1)^2}{2\sigma_1^2}\right) + (1-\phi)\frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(y_i-\mu_2)^2}{2\sigma_2^2}\right)\right)
\end{aligned}$$

Identification Challenge

Challenge of Maximum Likelihood Estimation

- The derivatives of the log-likelihood with respect to the parameters cannot be solved in closed form.
  - Setting the derivatives to zero does not yield straightforward solutions.
  - Solving for the maximum likelihood estimates analytically is not possible.
- The role of latent variables α_i:
  - Each α_i indicates the originating Gaussian of y_i.
  - If α_i's were known, the problem would be simplified.
  - Without knowing α_i, we must rely on iterative approaches like EM.

Try deriving the closed form solutions at home!

MLE for Gaussian Mixture I

- Derivative with respect to μ_1 and μ_2 Given the log-likelihood function for a Gaussian mixture model with two components, we derive the partial derivatives with respect to the means μ_1 and μ_2:

MLE for Gaussian Mixture II

- Derivative with respect to μ_1 The partial derivative of the log-likelihood with respect to μ_1 is:

$$\begin{aligned}\frac{\partial}{\partial \mu_1}\ell(\phi,\mu_1,\mu_2,\Sigma) &= \frac{\phi\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{(y_i-\mu_1)^2}{2\sigma_1^2}\right)}{\phi\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{(y_i-\mu_1)^2}{2\sigma_1^2}\right)+(1-\phi)\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left(-\frac{(y_i-\mu_2)^2}{2\sigma_2^2}\right)}\frac{(y_i-\mu_1)}{\sigma_1^2} \\ &= \underbrace{\frac{\phi f(y_i|\mu_1,\sigma_1^2)}{\phi f(y_i|\mu_1,\sigma_1^2)+(1-\phi)f(y_i|\mu_2,\sigma_2^2)}}_{w_{i1}}\frac{(y_i-\mu_1)}{\sigma_1^2}\end{aligned}$$

- Derivative with respect to μ_2 Similarly, the partial derivative of the log-likelihood with respect to μ_2 is:

$$\frac{\partial}{\partial \mu_2}\ell(\phi,\mu_1,\mu_2,\Sigma) = w_{i2}\frac{(y_i-\mu_2)}{\sigma_2^2}$$

MLE in Gaussian Mixtures

- The random variables α_i indicate which of the k Gaussians each y_i had come from.
- If we knew the values of α_i, the maximum likelihood problem would be easy.

$$\hat{\mu}_j = \sum_{i=1}^{n} \mathbb{1}\{\alpha_i = j\} y_i / \sum_{i=1}^{n} \mathbb{1}\{\alpha_i = j\} \qquad \text{for } j = 1, 2$$

$$\hat{\sigma}_j^2 = \sum_{i=1}^{n} \mathbb{1}\{\alpha_i = j\} (y_i - \hat{\mu}_j)^2 / \sum_{i=1}^{n} \mathbb{1}\{\alpha_i = j\} \qquad \text{for } j = 1, 2$$

$$\hat{\phi} = \sum_{i=1}^{n} \mathbb{1}\{\alpha_i = 1\} / n$$

- However, in density estimation, α_i's are not knownthis calls for algorithms like EM (Expectation-Maximization).

Understanding the EM Algorithm I

- **E-step (Expectation step):**
  - Start with $(\hat{\phi}, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2)$.
  - Calculate the posterior probabilities w_{ij} of the latent variables α_i given observations y_i and current parameter estimates.

$$w_{i1} = \frac{\phi f(y_i|\mu_1, \sigma_1^2)}{\phi f(y_i|\mu_1, \sigma_1^2) + (1-\phi) f(y_i|\mu_2, \sigma_2^2)}$$

  - $w_{i2} = 1 - w_{i1}$.
  - Each w_{ij} represents the probability that y_i was generated by the j-th component (Gaussian with mean μ_j and covariance σ_j^2).
  - Computationally, this involves evaluating the Gaussian density at y_i for each component and multiplying by the mixing probability ϕ_j.

Understanding the EM Algorithm II

- **M-step (Maximization step):**
  - Update the parameters μ_j, σ_j, and ϕ_j to maximize the likelihood based on the "soft" assignments w_{ij} from the E-step.
  - Updates are similar to those in a fully observed scenario but replace hard assignments with the weighted probabilities w_{ij}.

$$\hat{\mu}_j = \sum_{i=1}^{n} w_{ij} y_i / \sum_{i=1}^{n} w_{ij} \qquad \text{for } j = 1, 2$$

$$\hat{\sigma}_j^2 = \sum_{i=1}^{n} w_{ij} (y_i - \hat{\mu}_j)^2 / \sum_{i=1}^{n} w_{ij} \qquad \text{for } j = 1, 2$$

$$\hat{\phi} = \sum_{i=1}^{n} w_{i1} / n$$

Understanding the EM Algorithm III

- **Comparison with K-means:**
  - Unlike K-means, which uses hard clustering (assigns each point to a single cluster), EM uses soft clustering (points have probabilities associated with each cluster).
  - Susceptible to local optima; thus, multiple initializations might be necessary.

Understanding the EM Algorithm IV

- **Further Considerations:**
  - The EM algorithm iteratively refines estimates of both observed and latent variables.
  - Future discussions will expand on a generalized framework for EM, addressing other estimation problems with latent variables and providing proofs of convergence.

Applications of Mixture Models in Econometrics I

Mixture models are widely used in economic applications[1], such as:

- **Unobserved Heterogeneity in Labor and Industrial Organization (IO)**:
  - Berry, Carnall & Spiller (2006) Airline hubbing, costs and demand.
  - Keane & Wolpin (1997) The career decisions of young men. *Journal of Political Economy 105, 473522.*
  - Cameron & Heckman (1998) Life cycle schooling and dynamic selection bias: models and evidence for five cohorts. *Journal of Political Economy 106, 262311.*
- **Treatment of Multiple Equilibria in Discrete Games**:
  - Cooper (2002) Estimation and identification of structural parameters in multiple equilibria.
  - Berry & Tamer (2006) Identification in models of oligopoly entry.
  - Ciliberto & Tamer (2009) Market structure and multiple equilibria in airline markets. *Econometrica 77, 1791828.*

Applications of Mixture Models in Econometrics II

- **Measurement Error Models**:
  - Horowitz & Manski (1995) "Identification and robustness with contaminated and corrupted data."
  - Manski (2003) Partial identification of probability distributions.
- **Duration Models**:
  - Heckman & Singer (1984) A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica: Journal of the Econometric Society (1984): 271-320.*
  - Abbring & van den Berg (2003) "The nonparametric identification of treatment effects in duration models." *Econometrica 71.5 (2003): 1491-1517.*
- **Regime Switching Models**:
  - Cho & White (2007) Cho, Jin Seo, and Halbert White. *"Testing for regime switching." Econometrica 75.6 (2007): 1671-1720.*

[1] **Reference:** Compiani, G., & Kitamura, Y. (2016). Using mixtures in econometric models: a brief review and some new results. *Econometrics Journal*, 19(3), C95-C127.

Variations Finite Mixtures

Several variations of the finite mixture models.

1. **Multiple outcomes with independence properties are observed.**
2. Covariates enter the mixing weights.

Scenario 1: Multiple Outcomes with Independence Properties Are Observed I

In the statistical literature, Hall and Zhou (2003) consider two-component mixtures with independence property.

$$F(y) = \phi \sum_{k=1}^{K} F_{k,1}(y_k) + (1-\phi) \sum_{k=1}^{K} F_{k,2}(y_k)$$

where $y = (y_1, \ldots, y_k)^T$ and the cdf of each component factorizes by the independence assumption.

Understanding Clinical Trial Data

Motivation: The inspiration for this study originates from the clinical trial literature, where understanding and analyzing the data present unique challenges due to unobserved variables.

- **Vector of Outcomes**: y represents outcomes from k clinical tests. The distribution F varies depending on the disease status of a patient.
- **Observations**: Researchers observe the outcomes y, but the disease status of the patients and the proportion of affected individuals remain unobserved.
- **Statistical Challenge**: Nonparametrically identify and estimate the cumulative distribution functions (CDFs) $F_{k,j}$ and the mixture weight ϕ, utilizing random samples from the overall distribution F.

Scenario 2: Covariates Enter Mixing Weight

From Scenario 1, the model is, in general, not identified if the outcome is less than three-dimensional.

Henry, Kitamura, & Salanie (2013) improve the negative result by considering

$$F(y|x,w) = \sum_{j=1}^{J} \phi_j(x,w) F_j(y|x), \tag{2.2}$$

where y is now scalar-valued. Two main points about this model are:

1. Assuming data has **covariates satisfying certain exclusion restrictions**.
   - Covariate W affects the mixing weights, but does not change the component distributions.
2. Only **partially identified**. But the **characterisation of the identified set** enables **extracting useful information** from data.
   - E.g., nonparametric identification of the number of mixture components.

Markov Switching Models

Overview: Markov switching models assume that Y follows a finite-order autoregression conditionally on an unobserved Markov chain of order m.

- **Dependency:** The hidden Markov chain influences the outcome variable's distribution (e.g., expectation or variance).
- **Model Expression:** Can be formulated by setting $X = (Y_{t-1}, \ldots, Y_{t-m})$ and $W = (Y_{t-m-1}, \ldots, Y_1)$.

Misclassification Problem

Challenge: Researchers observe the outcome variable Y and potentially flawed measurements T of an underlying categorical regressor T^*.

- **Dependency:** While T and T^* are dependent, Y and T are assumed independent conditional on T^*.
- **Exclusion Variable:** Set excluded variable $W = T$, with $\phi_j = P\{T^* = t_j | T\}$.

Unobserved Heterogeneity

Context: Considering demand for a good with both observed and unobserved heterogeneity across buyers.

- **Geographical Variables (W):** Do not affect preferences directly but alter the distribution of buyer types.

Multiple Equilibria

Dynamics: Variables W do not influence the distribution of outcomes within an equilibrium but impact the equilibrium selection mechanism.

- **Case Study:** Ciliberto & Tamer (2009) - Anti-collusion policies do not affect a firms entry decisions but influence equilibrium selection across different regional markets.