Semi-Supervised Classification

Classification Maximum Likelihood Approach

Siwon Ryu

March 13, 2019

Outline

Introduction

Classfication Maximum Likelihood Criteria

Semi-supervised CEM-algorithm

Conclusion and future work

Introduction

Introduction

- Semi-supervised machine learning

- Supervised learning : Infers a function from labeled training data $\{(y_i, x_i) : i = 1, \cdots, n\}$
- Unsupervised learning : Infers a function from unlabeled data $\{x_i : i = 1, \cdots, n\}$
- Semi-supervised learning : Learn by using both labeled and unlabeled data $\{(y_1, x_1), \cdots, (y_n, x_n), x_{n+1}, \cdots, x_m\}$

Introduction

- Semi-supervised classification methods

- Co-training : [1] Blum et al.(1998)
- CEM-algorithm : [3] Celeux, Govaert (1992)
- Transductive Support Vector Machine : [10] Joachims, Thorsten (1999)
- Graph-Based Method : [5] Scholkopf et al. (2006)

Classfication Maximum Likelihood Criteria

Classfication Maximum Likelihood Criteria

- Clustering methods based on maximum likelihood

- Let $x = (x_1, \cdots, x_n)'$ be a given sample, $z_i = (z_{i1}, \cdots, z_{iK})$ be a vector of class indicators : $z_{ik} = 1$ if x_i is from class k and $z_{ik} = 0$ otherwise.
- Then, x is a sample from the following mixture densities(parametric)

$$f(x) = \sum_{k=1}^{K} \lambda_k f(x, \theta_k)$$

$\lambda_k \in (0, 1)$ are the mixing weights $(k = 1, \cdots, K)$, and $\sum_k \lambda_k = 1$.

- λ_k, θ_k can be chosen by maximizing following log-likelihood generally using EM-algorithm [8] Dempster (1977))

$$L = \log \prod_{i=1}^{n} \sum_{k=1}^{K} \lambda_k f(x_i, \theta_k)$$

Classfication Maximum Likelihood Criteria

- Clustering methods based on maximum likelihood : EM-algorithm

Suppose that $z = (z_1, \cdots, z_n)$ is unobserved, and the initial value of parameters $\theta^{(0)}$ is given.

- E-step

$$Q(\theta|\theta^{(q)}) = E_{z|x,\theta^{(q)}} \log L(\theta; x, z) = \sum_z \log L(\theta; x, z) P(z|x, \theta^{(q)})$$

- M-step

$$\theta^{(q+1)} = \arg\max_\theta Q(\theta|\theta^{(q)})$$

F.O.C.

$$\frac{\partial Q(\theta|\theta^{(q)})}{\partial \theta} = \sum_z \frac{\partial \log L(\theta; x, z)}{\partial \theta} P(z|x, \theta^{(q)}) = 0$$

Example) [11] Lee, Porter (1984)

$$
\begin{aligned}
\ln L(\theta, \lambda, p_{11}, p_{01}) = \sum_{t=1}^{T} & w_t \ln \left(f_1(y_t) p_{11} \lambda + f_2(y_t) p_{01} (1 - \lambda) \right) \\
& + (1 - w_t) \ln \left(f_1(y_t)(1 - p_{11}) \lambda + f_2(y_t)(1 - p_{01})(1 - \lambda) \right)
\end{aligned}
$$

- F.O.C. w.r.t. θ

$$
\frac{\partial \ln L}{\partial \theta} = \sum_{t=1}^{T} \left[P(1|y_t, w_t) \frac{\partial \ln f_1(y_t)}{\partial \theta} + P(0|y_t, w_t) \frac{\partial \ln f_2(y_t)}{\partial \theta} \right] = 0
$$

Classfication Maximum Likelihood Criteria

- Clustering methods based on maximum likelihood : EM-algorithm

Theorem. $L(\theta; x, z) = f(x, z|\theta)$ increases as $Q(\theta|\theta^{(q)})$ increases.

Proof.

Since $f(x, z|\theta) = f(x|\theta)P(z|x, \theta)$, we have

$$\begin{aligned}\log f(x|\theta) &= \sum_z P(z|x, \theta^{(q)}) \log f(x, z|\theta) - \sum_z P(z|x, \theta^{(q)}) \log P(z|x, \theta) \\ &= Q(\theta|\theta^{(q)}) + H(\theta|\theta^{(q)})\end{aligned}$$

and

$$\begin{aligned}\log f(x|\theta) - \log f(x|\theta^{(q)}) &= Q(\theta|\theta^{(q)}) - Q(\theta^{(q)}|\theta^{(q)}) + H(\theta|\theta^{(q)}) - H(\theta^{(q)}|\theta^{(q)}) \\ &\geq Q(\theta|\theta^{(q)}) - Q(\theta^{(q)}|\theta^{(q)})\end{aligned}$$

by Gibbs' inequality. $\square$

Classfication Maximum Likelihood Criteria

- Classification maximum likelihood approach

- Let $\lambda = (\lambda_1, \cdots, \lambda_K)$. The CML criterion* is defined by

$$C(z, \lambda, \theta) = \sum_{k=1}^{K} \sum_{i=1}^{n} z_{ik} \log \lambda_k f(x_i, \theta_k)$$

- In the CML approach, z, λ, θ are chosen by maximizing CML criterion.

Classfication Maximum Likelihood Criteria

- Classification maximum likelihood approach

- For example, when there are two classes,

$$C(z,\lambda,\theta) = \sum_{i=1}^{n} z_i \log \lambda_1 f(x_i,\theta_1) + (1-z_i)\log(1-\lambda_1) f(x_i,\theta_2)$$

F.O.Cs are

$$(\lambda_1) : \sum_{i=1}^{n} \left(\frac{z_i}{\lambda_1} - \frac{1-z_i}{1-\lambda_1} \right) = 0$$

$$\Rightarrow \hat{\lambda}_1 = \frac{1}{n} \sum_{i=1}^{n} z_i$$

$$(z_i) : \log \lambda_1 f(x_i,\theta_1) = \log(1-\lambda_1) f(x_i,\theta_2)$$

$$\Rightarrow \quad z_i = I(\lambda_1 f(x_i,\theta_1) > (1-\lambda_1) f(x_i,\theta_2))$$

$$(\theta_1) : \sum_{i=1}^{n} z_i \frac{\partial \log f(x_i,\theta_1)}{\partial \theta_1}$$

*[14] Scott, Symons (1971)

Classfication Maximum Likelihood Criteria

- Classification maximum likelihood approach : CEM-algorithm

Let $\theta^{(0)}, \lambda^{(0)}$ be given. Then, in the q^{th} iteration,

- E-step : Expectation

  Calculate posterior probabilities that x_i belongs to class k as

$$t_k^{(q)}(x_i) = \frac{\lambda_k^{(q)} f(x_i, \theta_k^{(q)})}{\sum_{k=1}^{K} \lambda_k^{(q)} f(x_i, \theta_k^{(q)})}$$

- C-step : Classification

  Assign each x_i to the cluster which provides the maximum $t_k^{(q)}(x_i)$:

$$z_{ik}^{(q)} = I\left(k = \inf\left\{m : t_m^{(q)}(x_i) \geq t_l^{(q)}(x_i) \quad \forall l = 1, \cdots, K\right\}\right)$$

- M-step : Maximization

  Calculate $\lambda_k^{(q+1)} = \frac{1}{n}\sum_{i=1}^{n} z_{ik}^{(q)}$, and obtain $\theta^{(q+1)}$ by maximizing CML criterion given $z_{ik}^{(q)}$, $\lambda_k^{(q+1)}$.

Classfication Maximum Likelihood Criteria

- Classification maximum likelihood approach : CEM-algorithm

Theorem.[†] Any sequence $\{z^{(q)}, \lambda^{(q)}, \theta^{(q)}\}$ of the CEM algorithem increases the CML criterion and the sequence $\left\{C(z^{(q)}, \lambda^{(q)}, \theta^{(q)})\right\}$ converges to a stationary value. Moreover, if the mixture estimates of the parameters are well-defined, the sequence $\{z^{(q)}, \lambda^{(q)}, \theta^{(q)}\}$ converges to a stationary position.

Proof.

- $C(z, \lambda, \theta)$ is increasing given z by M-step. i.e.,

$$C(z^{(q)}, \lambda^{(q+1)}, \theta^{(q+1)}) \geq C(z^{(q)}, \lambda^{(q)}, \theta^{(q)})$$

- We have also $C(z^{(q+1)}, \lambda^{(q+1)}, \theta^{(q+1)}) \geq C(z^{(q)}, \lambda^{(q+1)}, \theta^{(q+1)})$ since

$$\begin{aligned} z_{ik}^{(q+1)} = 1 \quad & \Leftrightarrow \quad t_k^{(q+1)}(x_i) \geq t_l^{(q+1)}(x_i) \quad \forall k \neq l \\ & \Leftrightarrow \quad \lambda_k^{(q+1)} f(x_i, \theta_k^{(q+1)}) \geq \lambda_l^{(q+1)} f(x_i, \theta_l^{(q+1)}) \\ & \Rightarrow \quad C\left(z^{(q+1)}, \lambda^{(q+1)}, \theta^{(q+1)}\right) \geq C\left(z^{(q)}, \lambda^{(q+1)}, \theta^{(q+1)}\right) \end{aligned}$$

- Since $K < \infty$, the increasing sequence $\{C(z^{(q)}, \lambda^{(q)}, \theta^{(q)})\}$ converges. Hence, if λ, θ are well-defined, the sequence $\{z^{(q)}, \lambda^{(q)}, \theta^{(q)}\}$ also converges.

† [3] Celeux, Govaert (1992) proposition 2

Semi-supervised CEM-algorithm

Semi-supervised CEM-algorithm

- CEM-algorithm using labeled data together with unlabeled data

- [13] Mclachlan (1992) extended CML-CEM algorithm to the case where both labeled and unlabeled data are used for learning.
- Let $x_l = \{(x_i, t_{ik}) : i = 1, \cdots, m\}$ be the labeled data, and $x_u = \{x_i : i = m+1, \cdots, n\}$ be the unlabeled data.
- The CML criterion in this case can be written as

$$L_c = \sum_{i=1}^{m} \sum_{k=1}^{K} t_{ik} \log f(x_i, \theta_k) + \sum_{i=m+1}^{n} \sum_{k=1}^{K} z_{ik} \log f(x_i, \theta_k)$$

L_c can be maximized by applying C-step to the unlabeled part.

Semi-supervised CEM-algorithm

- CEM-algorithm using misclassified label

- In practice, there are also classification error in the training data.
- Methods and result of Learning with imperfect labeled training data is proposed by [6] Chittineni (1980) [7] Chittineni (1981)
- Let $\hat{c}$ be the assigned class of x, and c is underlying true class of x.
- Density function when x_i belongs to class k is

$$
\begin{aligned}
f(x_i, \hat{c} = k) &= \sum_{l=1}^{K} f(x_i, \hat{c} = k, c = l) \\
&= \sum_{l=1}^{K} f(x_i | \hat{c} = k, c = l) P(\hat{c} = k, c = l)
\end{aligned}
$$

Semi-supervised CEM-algorithm

- CEM-algorithm using misclassified label

- Assume that the density of sample does not depends on its imperfect label given its true label :

$$f(x_i|\hat{c} = k, c = l) = f(x_i|c = l)$$

- Let $\alpha_{kl} = P(\hat{c} = k|c = l)$. Then, by Bayes rule,

$$f(x_i, \hat{c} = k) = p(x_i) \sum_{l=1}^{K} \alpha_{kl} P(c = l|x_i)$$

$$P(\hat{c} = k|x_i) = \sum_{l=1}^{K} \alpha_{kl} P(c = l|x_i)$$

Therefore, CML criterion is

$$L'_c = \sum_{i=1}^{m} \sum_{k=1}^{K} t_{ik} \log P(l = k|x_i) + \sum_{i=m+1}^{n} \sum_{k=1}^{K} z_{ik} \log \left(\sum_{l=1}^{K} \alpha_{kl} P(c = l|x_i) \right)$$

- Example : [11] Lee, Porter (1984)
  - Switching model :

$$y_t = x_t\beta + \delta I_t + \epsilon_t$$

  - Misclassified label : w

$$P(I_t = 1) = \lambda$$

$$P(w_t = 1|I_t = 1) = p_{11}$$

$$P(w_t = 1|I_t = 0) = p_{01}$$

Semi-supervised CEM-algorithm

- CEM-algorithm using misclassified label : Example

- CLM criterion

$$
\begin{aligned}
L'_c &= \sum_{t=1}^{T} z_i \log f(y_t, w_t, I_t = 1) + (1 - z_i) \log f(y_t, w_t, I_t = 0) \\
&= \sum_{t=1}^{T} z_i \log f_1(y_t)\,(w_t p_{11} + (1 - w_t)p_{10})\,\lambda \\
&\qquad + (1 - z_i) \log f_0(y_t)\,(w_t p_{01} + (1 - w_t)p_{00})\,(1 - \lambda)
\end{aligned}
$$

- Simulation result :

	n	100	500	1000	5000
CML	MSE	0.8113	0.4156	0.3626	0.3478
	Prob.mis.	0.0229	0.0139	0.0129	0.0128
ML	MSE	0.1773	0.0959	0.0635	0.0447
	Prob.mis.	0.0262	0.0216	0.0211	0.0208

$\lambda = 0.1,\ \delta = -5,\ p_{11} = 0.6,\ p_{01} = 0.4,\ \beta = (1, 0.7)',\ \sigma = 1, x = [1, N(0,1)]$

Conclusion and future work

Conclusion and future work

- Conclusion

Which one is better? Mixture approach or CML?

1. [4] Celeux, Govaert (1993)'s Simulation result : comparing CML vs. ML.
2. Symons (1981) : "There seems to be no simple recommendation to guide the users tof these criteria..."
3. [2] Bryant, Williamson (1978),[9] Ganesalingam (1989),... : CML criterion produces biased estimates of the mixture parameters. This bias can be tolerable if the mixture components are well separated and the proportions are not too extreme. ML is preferable.

Conclusion and future work

- Another methods to use imperfect information of sample separation

- Non-parametric supervised classification methods with imperfect training data
  1. Nearest neighbor
  2. Bayes classifier : assign class to maximize conditional probability.
- Noise-robust methods : [12] Liu, Tao (2016)
- EM-algorithm for nonparametric mixing distribution : [15] Train (2008)

References

-

[1] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training.
In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM, 1998.

[2] Peter Bryant and John A Williamson. Asymptotic behaviour of classification maximum likelihood estimates.
Biometrika, 65(2):273–281, 1978.

[3] Gilles Celeux and Gérard Govaert. A classification em algorithm for clustering and two stochastic versions.
Computational statistics & Data analysis, 14(3):315–332, 1992.

[4] Gilles Celeux and Gérard Govaert. Comparison of the mixture and the classification maximum likelihood in cluster analysis.
Journal of Statistical Computation and Simulation, 47(3-4):127–146, 1993.

[5] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews].
IEEE Transactions on Neural Networks, 20(3):542–542, 2009.

[6] CB Chittineni. Learning with imperfectly labeled patterns.
Pattern Recognition, 12(5):281–291, 1980.

References

-

[7] CB Chittineni. Estimation of probabilities of label imperfections and correction of mislabels.
Pattern Recognition, 13(3):257–268, 1981.

[8] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm.
Journal of the Royal Statistical Society: Series B (Methodological), 39(1):1–22, 1977.

[9] Selvanayagam Ganesalingam. Classification and mixture approaches to clustering via maximum likelihood.
Journal of the Royal Statistical Society: Series C (Applied Statistics), 38(3):455–466, 1989.

[10] Thorsten Joachims. Transductive inference for text classification using support vector machines.
In *Icml*, volume 99, pages 200–209, 1999.

[11] Lung-Fei Lee and Robert H Porter. Switching regression models with imperfect sample separation information–with an application on cartel stability.
Econometrica: Journal of the Econometric Society, pages 391–418, 1984.

[12] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting.
IEEE Transactions on pattern analysis and machine intelligence, 38(3):447–461, 2016.

[13] Geoffrey McLachlan. *Discriminant analysis and statistical pattern recognition*, volume 544.
John Wiley & Sons, 2004.

References

-

[14] Allen J Scott and Michael J Symons. Clustering methods based on likelihood ratio criteria. *Biometrics*, pages 387–397, 1971.

[15] Kenneth E Train. Em algorithms for nonparametric estimation of mixing distributions. *Journal of Choice Modelling*, 1(1):40–69, 2008.