S12 - Modelos de Topicos Latentes

Juan Carlos Martinez-Ovando *ITAM*

Mezclas Probabilisticas

Esta clase de modelos es ampliamente usada en la estadistica, *machine learning* y metodos semi- y no parametricos, generalmente. A su vez, es empleada profusamente para realizar **clasificacion no supervisada** de datos, con el proposito de revelar *agrupaciones subyacentes* en datos.

Aun cuando estos modelos son empleados para realizar **clasificacion supervisada**, su origen es el de esimacion de densidades.

Pensemos que X es una variable aleatoria (absolutamente) continua, con funcion de densidad $f(x)$. En vez de emplear o comprometerse con solo una funcion de densidad (parametrica), el modelo contempla que la densidad de X puede describirse como una combinacion lineal convexa de multiples funciones de densidades, i.e.

$$f(x) = \sum_k w_k f(x|\theta_k),$$

donde las $f(\cdot|\theta_k)$s son funciones parametricas de densidades, las cuales difieren solo en terminos de los diferentes valores de los parametros θ_ks, y los pesos de la mezcla w_ks definen una combinacion lineal convexa de las $f(\cdot|\theta_k$s.

La combinacion lineal convexa anterior es bastante flexible, pues puededefinirse de manera *densa* en la clase de **todas** las distribuciones absolutamente continuas con soporte en $\mathcal{X}$ (e.g. densidades multimodales, sesgadas, simetricas, etc.).

¿Cual es la relacion con metodos de clasificacion no supervisada?

Bueno, pues un resultado *muy circunstancial*, para efectos inferenciales, permite extender el modelo con la inclusion de **variables latentes**, z, que permiten indicar de que compoente entres las $f(|\theta_k)$ la variable X es generada. Asi, la expresion extendida del modelo resulta en,

$$f(x,k) = P(z=k) \times P(x|z=k) = w_k \times f(x|\theta_k) \times 1(z=k),$$

siendo entonces las w_ks entendidas como las probabilidades (de un procedimiento multinomial excluyente) de que la variable x sea descrita por el componente $f(x|\theta_k)$.

Para un conjunto de datos, $x_1, \ldots, x_n$, se sigue entonces que la verosimilitud (extendida) incluyendo las variables latentes, $z_1, \ldots, z_n$, esta dada por

$$lik(w, \theta, z|x) = \prod_i w_{z_i} f(x_i|\theta_{z_i}) = \prod_k w_k^{\#\{z_i=k\}} f(\{x_i : z_i = k\}|\theta_k),$$

por el componente multinomial.

Asi, el procedimiento extendido da origen a un **procedimiento circunstancial** de clusterizacion. Es un metodo bastante flexible, pues la clasificacion no supervisada descansa en argumentos probabilisticos y no en una nocion de distancia (como otros metodos no supervisados de clasificacion).

Lo anterior da origen a que podamos extender la nocion de mezclas probabilisticas a contextos donde las variables no sean (absolutamente) continias, sino *discretas* y/o *categoricas*, entre otras.

Latent Dirichlet Allocation (LDA)

El modelo LDA es un procedimiento de clasificacion ni supervisada de contenido de textos, cuya clasificacion resultante es entendida como la *revelacion de topicos latentes*.

Para este efecto, como hemos comentado antes, pensemos que un conjunto de textos, $t_1, \ldots, t_n$ esta referido a un **diccionario lexico** con D palabras relevantes (no ordenadas). Cada texto es codificado vectorialmente como el vector de frecuencia de palabras en el diccionario lexico que aparecen en el mismo, i.e.

$$t_i \approx x_i,$$

donde $x_i \in \mathbb{N}^D$ donde x_{id} es el numero de veces que la palabra d del diccionario lexico aparece en el texto t_i, para $d = 1, \ldots, D$.

De esta forma podemos pensar que la frecuencia de palabras de cada texto puede describirse con ladistribucion multinomial,

$$x_i \sim Mult(x_i|N_i, \theta_1, \ldots, \theta_D),$$

donde N_i es el numero de palabras en el texto i y las θ_ds son las probabilidades de que la palabra d del diccionario lexico aparezca en el texto.

Topicos latentes

Los topicos latentes de un conjunto de textos bajo LDA pueden asociarse con diferentes frecuencias/repeticiones de palabras o terminos, caracterizados a su vez por diferentes θ_ds bajo la representacion multinomial.

Asi, si pensamos que puede haber K posibles topicos latentes, podremos pensar en K posibles configuraciones de $(\theta_{ik}, \ldots, \theta_{Dk})_{k=1}^K$ asociadas.

De esta forma, adaptando el modelo de mezclas probabilisticas tenemos que la incertidumbre sobre el contenido de un texto puede describirse como

$$p(x_i) = \sum_k w_k Mult(x_i|N_i, \theta_{1,k}, \ldots, \theta_{D,k}),$$

interpretando las w_ks como antes.

Extendiendo a la inclusion de la variables asignacion latente, z_i tenemos

$$p(x_i, z_i) = w_{z_i} Mult(x_i|N_i, \theta_{1,z_i}, \ldots, \theta_{D,z_i}).$$

El aprendizaje o inferencia estadistica en esta clase de modelos es bastante compleja, pues los calculos no pueden obtenerse de manera analitica cerrada.

Bajo el paradigma bayesiano de inferencia, la estimación de los parametros y variables latentes descansan tipicamente en metodos numericos de simulacion basados en MCMC. Sin embargo, estos algoritmos son costosos computacionalmente y no escalables.

En la actualidad, una alternativa para resolver la limitante anterior descansa hace uso de **aproximaciones variacionales**, que brevemente describimos a continuacion.

Variational Bayes

En el modelo anterior, las variables $(w, \theta) = (w_k, \theta_k)_{k \geq 1}$ definen el conjunto de parmaetros, mientras que $z = (z_j)_{j=1}^n$ denota el conjunto de variables latentes. Inferencia sobre esta clase de modelos se basa en a distribucion final,

$$p(w, \theta, z|x) \propto p(x|w, \theta, z) p(z|w, \theta) p(w, \theta).$$

Como mencionamos, la idea de los metodos variacionales consiste en aproximar $p(w, \theta, z|x)$ por una funcion $q(w, \theta, z)$ de manera que

$$p(x) = p(q) + KL(q, p),$$

siendo $KL(q, p)$ la divergencia de Kullback-Leibler entre p y q, i.e.

$$KL(q, p) = -\int \log\left(\frac{p(w, \theta, z|x)}{q(w, \theta, z)}\right) Q(dw, d\theta, dz).$$

La idea es que $KL(q, p)$ sea pequeña. De toda forma, $\tilde{p} = \exp\{p(q)\}$ es una cota inferior de $p(x)$.

Variational Bayes descansa sobre el procedimiento MAP (Maximum a Posteriori) como alternativa del enfoque general bayesiano. (Esto es bien justificado en terminos de teoria de la decision).

Asi, en vez de maximizar $p(w, \theta, z|x)$ el algoritmo maximiza $q(w, \theta, z)$, por medio de minimizar $KL(q, p)$.

El algoritmo adapta justamente $q(w, \theta, z)$, por lo que en la practica nunca alcanza a empatar q con p. De esta forma, el procedimiento es aproximado. Mas aun, pues para acelerar los calculos computacionales, la eleccion de $q(w, \theta, z)$ se restringe a una clase de distribuciones manejables.

Ilustacion

Paquetes

Empleamos en estas notas dos paquetes: `RTestTools`, empleado solo para recuperar los atos para la ilustracion de los procedimientos, y `topicmodels`, por la implementacion del algoritmo variacional para LDA. Esta ilustracion fue realizada en `MRO 3.4.4`.

```
if(!require('RTextTools')){install.packages("RTextTools")}
if(!require('topicmodels')){install.packages("topicmodels")}
```

Datos

```
library("RTextTools")
data(NYTimes)
data <- NYTimes[ sample(1:3100, size=1000,replace=F), ]
dim(data)
```

```
## [1] 1000    5
```

```
head(data)
```

```
##      Article_ID      Date
## 1518      27908  7-Mar-01
## 1929      20338 22-Jul-02
## 647       31507 18-Mar-98
## 334       39467 18-Feb-97
## 682       31857 28-Apr-98
```

```
## 206       43575 12-Sep-96
##                                                                  Title
## 1518        Senate Votes to Repeal Rules Clinton Set on Work Injuries
## 1929 More Say Yes to Foreign Service, But Not to Hardship Assignments
## 647                            Chase Will Lay Off 2,250 in Latest Cuts
## 334  U.S. to Pay New York Hospitals Not to Train Doctors, Easing Glut
## 682       Restrictions on Iraq Will Stay in Force, U.N. Council Rules
## 206      POLITICS: THE MONEY; A Hollywood Production: Political Money
##                                                                Subject
## 1518                          Senate votes to repeal workplace injuries rules
## 1929                                       hiring of Foreign Service officers
## 647  Chase Manhattan, nation's largest bank plans to lay off 2,250 employees
## 334     Federal government to pay New York hospitals not to train physicians
## 682              U.N. Security Council votes to extend sanctions against Iraq
## 206                                          clinton fundraising in hollywood
##      Topic.Code
## 1518          5
## 1929         19
## 647          15
## 334           3
## 682          16
## 206          20
```

```
matrix <- create_matrix(cbind(as.vector(data$Title),
                              as.vector(data$Subject)),
                        language="english",
                        removeNumbers=TRUE,
                        stemWords=TRUE)
```

```
k <- length(unique(data$Topic.Code))
```

Implementacion del algoritmo

```
library("topicmodels")
lda.out <- LDA(matrix, k)
summary(lda.out)

print(lda.out@gamma[1,])
```

Referencias adicionales

- **Jordan**, Graphical Models, *Statistical Science*
- **Titterington**, "Bayesian Methods for Neural Networks and Related Modelos", *Statistical Science*
- **Bishop**, *Pattern Recognition and Machine Learning (Book)*

- **Minka & Winn**, `infer.NET`, Microsoft Research, link