Modern Computational Statistics lec01 Notes

Author:Cao Yihan

This is based on PKU course: Modern Computational Statistics. Thanks to Prof. Cheng Zhang, this is a very interesting course.

1 Statistical Pipeline

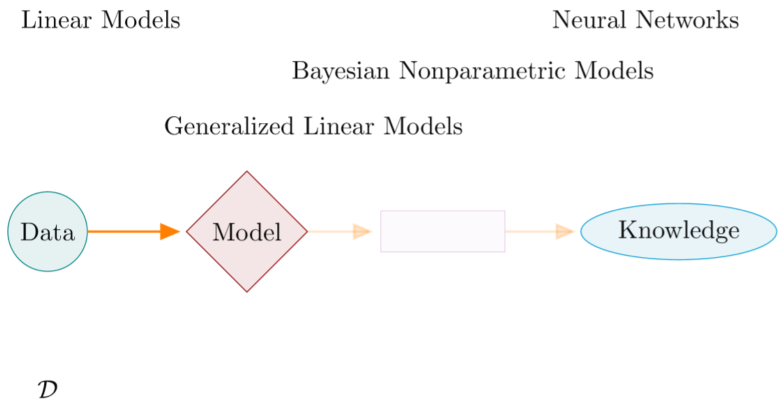

Figure 1: Statistical Pipeline

This graph described a general statistical pipeline from data to Knowledge. According to the data we have, we can build the model. In this picture, it lists some of those common models we like to use, like linear models,generalized linear models, bayesian nonparametric models and neural networks. Models always contain parameters. Based on models, next step is to estimate the parameters.

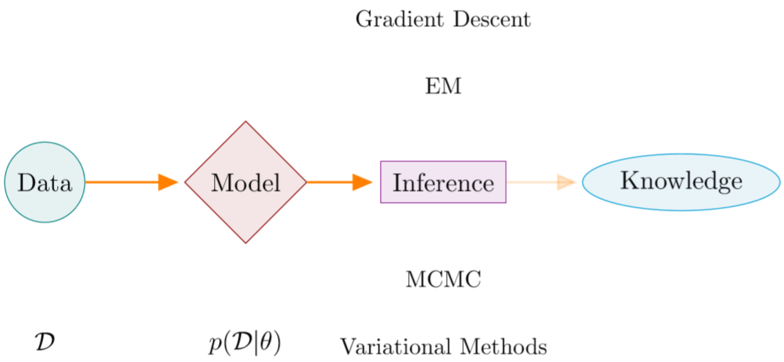

Figure 2: Statistical Pipeline

This graph described some inferences ways for estimation. In computational statistics, we have methods like gradient descent, EM and MCMC algorithm, and some modern variational methods. These will be covered in the upcomming courses. After the inference step,we can finnaly get the estimation of parameters in model $p(D|\theta)$.

Here in computational statistics, we mainly focus on model and inference.

2 Linear Models

Suppose we have data: $D = \{(x_i, y_i)\}_{i=1}^n$. Build a linear model based on this data:

$$Y = X\theta + \epsilon, \epsilon \sim N(0, \sigma^2 I_n) \rightarrow Y \sim N(X\theta, \sigma^2 I_n)$$

2.1 One example here:Gaussian Mixture Model

Gaussian mixture models are a probabilistic model for representing normally distributed subpopulations within an overall population. In a large dataset,D, mixture models don't require knowing which subpopulation a sample belongs to, allowing the model to learning subpopulations automatically.

Dataset D always has more than one "peak", if it belongs to a gaussian mixture population. Using a single gaussian model gives a poor fit. This is when we should use GMM model.

Recalling a multi-dimensional model:

$$p(x) = \sum_{i=1}^{K} \Phi_i N(x|\mu_i, \Sigma_i)$$

$$N(x|\mu_i, \Sigma_i) = \frac{1}{\sqrt{(2\pi)^K |\Sigma_i|}} exp(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i))$$

$$\sum_{i=1}^{K} \Phi_i = 1$$

Here Φ_i represents the weights. And GMM is the sum of several multidimensional models:

$$p(x) = \sum_{k=1}^{K} p(k)p(x|k) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)$$

where $\sum_{i=1}^{K} \pi_i = 1$ represents the weights.

The commom method of estimating the parameters is maximum likelihood estimation. First, take the log of the objection function.

$$max \sum_{i=1}^{N} logp(x_i) \rightarrow max \sum_{i=1}^{N} log(\sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k))$$

However, it's analytically impossible for differentiating the log likelihood and solving for 0. So here we should use EM algorithm for convergence, when we estimating the parameter, which will be covered later in this course.

2.2 One example: Latent Dirichlet Allocation

Latent Dirichlet Allocation is a basic model for topic modeling in learning latent preferences. It's uses can also be expanded to recommender systems. Below is the commom graph for LDA, which is often seen in papers.

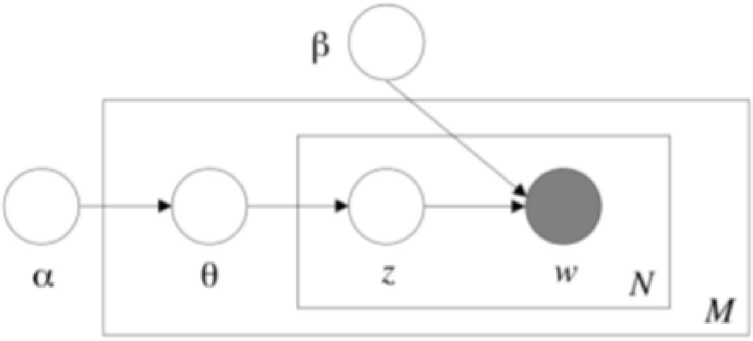

Figure 3: LDA model

Suppoe the dataset is: $D = \{w_i\}_{i=1}^{M}$. Now we have a mixture of topics $\theta \sim Dir(\alpha)$(α is a parameter in Dir). For each of the words(N) w_n, we suppose:

$$z_n \sim Multinomial(\theta), w_n|z_n, \beta \sim p(w_n|z_n, \beta)$$

And using bayesian inference, we have:

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn},\beta) d\theta_d$$

The graph describes this process. See more in the paper: LDA matching.

3 Exponential Families, Score Function and Fisher Information

3.1 Exponential Families

Exponential family distributions always take the following form:

$$p(y|\theta) = h(y)exp(\phi(\theta)T(y) - A(\theta))$$

$$A(\theta) := log(\int_y h(y)exp(\phi(\theta)T(y))dy), whenever \quad A \quad is \quad finite.$$

where $\phi(\theta)$ is natural parameters and $T(y)$ is sufficient statistics. To express in a more formal way:

$$p_\theta(x) = h(x)exp(< \theta, \phi(x) > -A(\theta))$$

It's obvious that Bernoulli distribution, poission distribution and normal distribution all belong to exponential families. There are many reasons to study exponential families. They arise as the solutions to several natural optimization problems. They also enjoy certain robustness properties related to optimal Bayes' procedures. They are analytically very tractable and have been the objects of substantial study for nearly the past hundred years. See more in Information Theory part. [1]

3.2 Score Function

Score function is the partial derivative of the log-likelihood function, where is the standard likelihood function.[1]

$$s(\theta) = \frac{\partial L}{\partial \theta} \quad whereL(\theta;Y) = \sum_{i=1}^{n} logp(y_i|\theta)$$

The expected value of the score is zero.

[1] http://mathworld.wolfram.com/ScoreFunction.html

3.3 Fisher Information

Fisher information can be seen as the variance of the score function. In a formal way:

$$I_\theta := E_\theta[\nabla_\theta logp_\theta(X)\nabla logp_\theta(X)^T] = E_\theta[l_\theta l_\theta^T]$$

Under mild assumptions, like for exponential families,

$$I(\theta) = -E(\frac{\partial^2 L}{\partial\theta\partial\theta^T})$$

Fisher information is a measure of the curvature of Log-likelihood function. It reflects the sensitivity of model about the parameter at its current value. Fisher information captures the variability of gradient $\nabla logp_\theta$.

Here we should also mention KL divergence. The full name for KL divergence is Kullback-Leibler divergence. It is a measure of statistical distance between two distributions.

$$D_{KL}(q||p) = \int q(x)log\frac{q(x)}{p(x)}dx$$

It's easy to prove that KL divergence is non-negative.

This also gives a rigorous definition of mutual information. [2]If X and Y are random variables with joint distribution P_{XY} and marginal distributions P_X and P_Y, define:

$$I(X;Y) = D_{kl}(P_{XY}||P_X \times P_Y)$$

Notice that Fisher information is Hessian of KL-divergence between two distributions $p(x|\theta)$ and $p(x|\theta')$ when $\theta' = \theta$.

$$(HessianMatrix) \quad \nabla^2_{\theta'} D_{KL}(p(x|\theta)||p(x|\theta'))|_{\theta'=\theta} = I(\theta)$$

There are also some contents for MLE, Bayesian inference and markov chains. Because these knowledge are a little basic, they are not mentioned here. See the original slide for more information.

References

[1] Lecture Notes for Statistics 311/ EE 377, by John Duchi,Stanford,Chapter14.

[2] Lecture Notes for Statistics 311/ EE 377, by John Duchi,Stanford,Page 19.