Lecture 11 : Smoothing: Penalized Likelihood Methods and GAMs

Tyrell Richu

10/23/2018

Penalized Likelihood

- Consider an arbitrary model with generic parameter β, a log-likelihood function L(β).

- Let λ(.) denotes the roughness penalty which decreases as the values of β are smoother (i.e. uniformly close to zero). The penalized likelihood estimator L*(β) is :

$$L^*(\beta) = L(\beta) - \lambda(\beta),$$

- Penalized likelihood methods are examples of regularization methods. It is a general approach for modifying ML methods to give sensible answers in unstable situations such as modeling using data sets consisting too many variables.

Types of Penalties $\lambda(\beta)$

- L_2-norms (Ridge Regression) : $\lambda(\beta) = \lambda \sum_j \beta_j^2$
- L_1-norms (LASSO) : $\lambda(\beta) = \lambda \sum_j |\beta|$, subject to the constraint $\sum_j |\beta| \leq$ K, where K is some constant.
- L_0-norms : $\lambda(\beta) \propto non - zero \beta_j$ -AIC/BIC methods are a special case of L_{0}-penalization but it's hard to optimize for large j.

How to select $\lambda(\beta)$ for penalized likelihood

-The degree of smoothing depends on the smoothing parameter λ, the choice of which reflects the bias/variance trade-off. When λ increases, the estimates $\{\beta_j\}$ decrease towards zero, thus decreasing the variance but increases the bias.

- K-fold Cross-validation Goal : We are interested in choosing a λ based on fitting the model to part of the data and then checking the goodness of fit in terms of prediction for the remaining data.
- Step 1: Fix λ'.
- Step 2: Do this k-times, leave out the fraction 1/k of the data and predict it using the model fit for the remaining data. Choose the value of λ which has the lowest prediction error.
- Step 3: Compute the error for λ'
- Step 4: Repeat for k-values of λ. Then, choose the value of λ which has the lowest prediction error.

Note: Bayesian methods can also approximate penalized likelihood if $prior(\beta) \propto exp(-\lambda(\beta)) = posterior(\beta) \propto L^*(\beta)$

Pros/Cons of Penalized Likelihood

- L_2-norms (-) : Useless for finding a rigid model, because all the variables remain in the model.
- L_1-norms (+) : Allows us to plot estimates as a function of λ to summarize how explanatory variables, β_j drop out as λ increases by selecting individual indicators rather than entire factors.
- L_1-norms (-) : May overly penalize β_j that are truly large may hold high bias, making inference difficult. Solution: adjust the penalty function such that it includes both the L_{1} and L_{2} norms.

General Additive Models (GAMs)

- GAMs are another type of GLM that specifies a link function g(.) and a distribution for the random component.
- In GLMs, we had g(μ_i) = $\sum_j \beta_j x_{ij}$
- In GAMs, g(μ_i) = $\sum_j s_j(x_{ij})$, where s_{j}(.) is unspecified smooth function of predictor j. Examples: cubic splines: cubic polynomials over sets of disjoint intervals, joined together at boundaries called knots.

-We can fit GAMs using the backfitting algorithm, similar to Newton's method, to utilize local smoothing.

- Step 1: Initialize $s_j = 0$
- Step 2: For each rth iteration, update s_j such that

$$s_j^{(r)} = y_i^{(r)} - \sum_{k \neq j} s_k^{(r)}(x_{ik}), j = 1, ..., p$$

- This will fit a model that assigns a deviance and an approximate degree of freedom to each s_j in the additive predictor, allowing inference about each term. The df helps determine how smooth the GAM fit looks. (e.g. Smooth functions with df = 4 look similar to cubic polynomials, which has 4 parameters)
- Like with GLMs, we can compare deviances for nested models to test whether a model gives a significantly better fit than a simpler model.

Final Notes

- GAMs and penalized likelihood methods are stronger than GLMs because they impersonate GLMs in assuming a binomial distribution for a binary response and having a df value associated with each explanatory effect.