PH240C LAB 01

Sep 14, 2021

Agenda:

- About the labs
- Final project policy
- Review of lectures + More on GLM
- Computation aspects
- Homework 1 hints

Berkeley
UNIVERSITY OF CALIFORNIA

About the labs…

- GSI: Lei Shi, 2nd year Ph.D. in Biostats
- Schedule: Biweekly
- Mode: Hybrid + Recorded, but encourage in-person participation
- What will we do in labs?
  - Review lectures and walk little bit further
  - More interesting topics and examples
  - Homework hints!

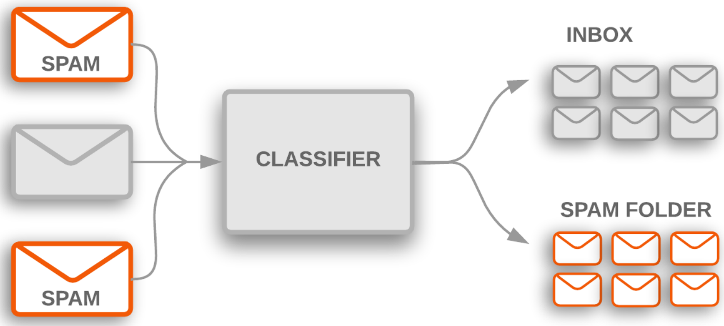
Final project policy

- Final project write-up and presentation takes up 35% of the final grade
- Presentation date: Dec 08
- Teams: 1 or 2 persons. Larger group size needs permission.
- Topic: Analyze data using modern statistical learning algorithms.
  - Highly recommend: finding your own data(from your projects or asking companies for help) If not possible you could use our assigned data too.

Berkeley
UNIVERSITY OF CALIFORNIA

Review of the lectures

- Classification: We have data (X_i, Y_i) from some unknown distribution F. $Y_i \in \{0,1\}$ are binary.



- We hope to find a classifier $h \in \mathscr{H}$ to minimize the generalization risk

$$R(h, F) = E_F(1\{Y \neq h(X)\}).$$

Berkeley
UNIVERSITY OF CALIFORNIA

Review of the lectures

- Logistic regression:

$$E(Y \mid X) = \mu(X) \in [0,1], Y \mid X \sim \text{Bernoulli}(\mu(X)).$$

Choosing $\mu(X) = \text{expit}(X'\beta) = \dfrac{\exp(X'\beta)}{1 + \exp(X'\beta)}$ gives logistic regression.

- Given the data, We obtain an estimator $\hat{\beta}$ with maximum likelihood estimation
- The classifier is then given by

$$y = \begin{cases} 1 & \text{if } \text{expit}(x'\widehat{\beta}) \geq 0.5 \quad \Leftrightarrow \quad x'\widehat{\beta} \geq 0 \\ 0 & \text{if } \text{expit}(x'\widehat{\beta}) < 0.5 \quad \Leftrightarrow \quad x'\widehat{\beta} < 0. \end{cases}$$

Review of the lectures

- Support vector machines: find hyperplane $\mathscr{A}$ that separates the training data and maximize the minimal margin
- Mathematically we solve optimization:

$$\min_{a,\mathrm{w}} \quad \mathrm{w}'\mathrm{w}$$
$$\text{s.t.} \quad Y_i(a + \mathrm{w}'X_i) \geq 1, \quad i = 1, \ldots, n.$$

- In many cases we cannot find or deliberately avoid separating plane by introducing soft constraints:

$$\min_{a,\mathrm{w}} \quad \mathrm{w}'\mathrm{w} + C\sum_{i=1}^{n} \xi_i$$
$$\text{s.t.} \quad Y_i(a + \mathrm{w}'X_i) \geq 1 - \xi_i,$$
$$\xi_i \geq 0, \quad i = 1, \ldots, n.$$

Berkeley
UNIVERSITY OF CALIFORNIA

More on logistic regression

- If we have high dimensional covariates, we want a sparse $\hat{\beta}$ (a parsimonious and more interpretable model)
- Example: use gene expression levels(thousands of covariates) to predict certain disease.
- LASSO on GLM: Adding ℓ_1 penalty on the log likelihood:

$$\widehat{\beta} = \arg\max_{b \in \mathbb{R}^d} \log L_n\left(b; \{(Y_i, X_i)\}_{i=1}^n\right) + \lambda \|b\|_1$$

Computation aspects

- Logistic regression can be solved with Newton-Raphson.
- SVM can be solved by quadratic programming(QP).
    - A Kernel trick can be applied to adapt to nonlinear classifiers. Here's a link for SVM in R.
- We have existing R packages for GLM(stats, glmnet) and SVM(e1071).
- See our R code file: glm_svm.r.
- Test classification reference:
    1. Text classification with tidy data principles
    2. Practicing sentiment analysis with Harry Potter

Homework 1 hints

- Problem 1: Predicting GPA
  - Recall how to predict a success probability with a logistic model
  - Here "odds" means odds ratios!
  - Build an equation and solve for X_1. Please give the numerical result(i.e. no logs or exps).

Berkeley
UNIVERSITY OF CALIFORNIA

Homework 1 hints

- Problem 2: SVM
  - Graph paper not required. Three regions related to a hyperplane. How to determine the signs:
    - A dumb try always works
    - Geometric interpretation
  - Recall what is margin. Consider how to calculate distance between paralleled planes.
  - Use the hyperplane!
  - Two ways of calculating slack variable here:
    - A hinge loss interpretation from wikipedia: SVM
    - For slack variables the inequalities in the constraints can be attained

Homework 1 hints

- Problem 3: Heart disease data
  - SVM functions and different kernels
  - Try your hand!