Topic 2: Classification

Eric B. Laber

Department of Statistical Science, Duke University

Statistics 561

On choices

I'd rather be rich than stupid.
—Pierre-Joseph Proudhon

On having to cover classification in two lectures

I needed to think last night. So I galloped into a wooded glen, and after punch dancing out my rage and suffering an extremely long and very painful fall, I realized what has to be done.
*—Rod*

Warm-up (5 minutes)

- Explain to your group
    - What is classification? What are come canonical examples?
    - What is the separability problem?
    - What is zero-one loss?

- True or false
    - Regression is a misnomer in 'logistic regression'
    - Classification is in the "inner loop" of many RL algorithms
    - McDonald's is the most prevalent fast-food chain in NC

Classification: quick overview

- Input-output pairs where output is one of finitely many categories
  - High-risk v low-risk for complications in surgery
  - Spam v not spam
  - Handwritten digit recognition
- Example from sex-trafficking classification problem
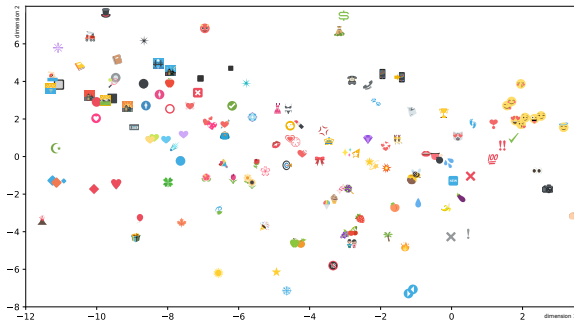
Setup: binary classification

▶ Observe $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^{n}$ comprising i.i.d. draws from P

  ▶ Inputs: $\boldsymbol{X} \in \mathbb{R}^p$

  ▶ Outputs: $Y \in \{-1, 1\}$, aka, label

▶ Classifier $c : \mathbb{R}^p \to \{-1, 1\}$ so that $c(\boldsymbol{x})$ is the predicted label at input $\boldsymbol{X} = \boldsymbol{x}$

0-1 loss

▶ Natural measure of classification performance is 0-1 loss

$$\ell_0(\boldsymbol{x}, y; c) \triangleq 1_{y \neq c(\boldsymbol{x})}$$

so that the expected loss (risk) is

$$\tau(c) \triangleq P\ell_0(\boldsymbol{X}, Y, c) = P1_{Y \neq c(\boldsymbol{X})} = P\{Y \neq c(\boldsymbol{X})\}$$

▶ E.g., linear classifier $c(\boldsymbol{x}; \boldsymbol{\beta}) = \mathrm{sign}(\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta})$ which has loss

$$\tau(\boldsymbol{\beta}) \triangleq P1_{Y \neq \mathrm{sign}(\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta})} = P1_{Y\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta} < 0}$$

Bayes classifier

▶ Let $\mathcal{C}$ be class of all (msbl) maps from $\mathbb{R}^p$ into $\{-1, 1\}$ and

$$c^{\mathrm{opt}} = \arg\min_{c \in \mathcal{C}} \tau(c) = \arg\min_{c \in \mathcal{C}} P\{Y \neq c(\boldsymbol{X})\}$$

then

$$c^{\mathrm{opt}}(\boldsymbol{x}) = \left\{ \begin{array}{cc} 1 & \text{if } P(Y = 1 | \boldsymbol{X} = \boldsymbol{x}) \geq 1/2 \\ -1 & \text{otherwise} \end{array} \right.$$

i.e., $c^{\mathrm{opt}}(\boldsymbol{x}) = \mathrm{sign}\{2q(\boldsymbol{x}) - 1\}$ w/ $q(\boldsymbol{x}) \triangleq P(Y = 1 | \boldsymbol{X} = \boldsymbol{x})$

Blank page for notes

Blank page for notes

Probabilistic classifiers

▶ Natural approach to classification is to estimate construct an estimator $\widehat{q}_n(\boldsymbol{x})$ of $q(\boldsymbol{x}) = P(Y = 1|\boldsymbol{X} = \boldsymbol{x})$

▶ Logistic regression posits a model of the form

$$q(\boldsymbol{x}; \boldsymbol{\beta}^*) = \operatorname{expit}(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta}^*) = \frac{\exp(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta}^*)}{1 + \exp(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta}^*)}$$

Derive log-likelihood for logistic regression

Blank page for notes

In-class exercise (5 min)

▶ In your group derive a stochastic gradient descent algorithm
  for logistic regression: then on to logistic_regression.R

Separability problem

▶ If classes linearly separable estimates diverge

▶ Use ridge with logistic regression[1]

▶ Penalized negative log-likelihood (0-1 coding for simplicity)

$$\ell_n(\boldsymbol{\beta}) = -\mathbb{P}_n\left[\boldsymbol{X}^{\mathsf{T}}\boldsymbol{\beta}Y + \{1 - q(\boldsymbol{X};\boldsymbol{\beta})\}\right] + \lambda||\boldsymbol{\beta}||^2$$

differentiate wrt $\boldsymbol{\beta}$ to obtain

$$-\mathbb{P}_n\boldsymbol{X}\left\{Y - q(\boldsymbol{X};\boldsymbol{\beta})\right\} + 2\lambda\boldsymbol{\beta}$$

---

[1]That's right, I'm stating this without qualification of any kind. What's it to you? That's what I thought.

Large-margin classifiers

▶ Decision boundary $\{\boldsymbol{x} : q(\boldsymbol{x}) = 1/2\}$

▶ Don't need $P(Y|\boldsymbol{X} = \boldsymbol{x})$ just $\text{sign}\{2q(\boldsymbol{x}) - 1\}$, i.e., we don't need to know the boundary just which side we're on

▶ Idea! Construct loss functions that penalize distance from correct side of boundary

Surrogate loss functions: laber draws a picture

Surrogate loss functions

▶ Common surrogate loss functions include

  ▶ Squared error loss: $\ell(\boldsymbol{x}, y; \boldsymbol{\beta}) = (1 - y\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta})^2$

  ▶ Exponential loss: $\ell(\boldsymbol{x}, y; \boldsymbol{\beta}) = \exp(-y\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta})$

  ▶ Hinge loss: $\ell(\boldsymbol{x}, y; \boldsymbol{\beta}) = (1 - y\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta})_+$

  ▶ Logistic loss: $\ell(\boldsymbol{x}, y : \boldsymbol{\beta}) = \log\{1 + \exp(-y\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta})\}$

▶ Can also consider penalized versions of these loss functions

Nice and relaxing derivation

▶ Show $\widehat{\boldsymbol{\beta}}_n = \arg \min\limits_{\boldsymbol{\beta}} \mathbb{P}_n \log \{1 + \exp(-Y\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta})\}$ recovers the logistic regression estimator

Blank page for notes

Statistical inference

- ▶ Convexity buys you a lot, many derivations much simpler

- ▶ High-level overview of assumptions

  (A1) $\ell(\boldsymbol{x}, y; \boldsymbol{\beta})$ is convex in $\boldsymbol{\beta}$ for each $(\boldsymbol{x}, y)$

  (A2) $Q(\boldsymbol{\beta}) = P\ell(\boldsymbol{X}, Y; \boldsymbol{\beta})$ exists and is finite for all $\boldsymbol{\beta}$

  (A3) $\boldsymbol{\beta}^* = \arg\min Q(\boldsymbol{\beta})$ exists an is unique

  (A4) $Q(\boldsymbol{\beta})$ is twice continuously differentiable in a nbrhd of $\boldsymbol{\beta}^*$ and $H = \nabla^2 Q(\boldsymbol{\beta}^*)$ is positive definite

Theorem
Assume (A1)-(A4) and let $\Omega = P\nabla\ell(\boldsymbol{X}, Y; \boldsymbol{\beta}^*)\nabla\ell(\boldsymbol{X}, Y; \boldsymbol{\beta}^*)^\intercal$ then

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) \rightsquigarrow \mathrm{Normal}(0, H^{-1}\Omega H^{-1}).$$

Statistical inference cont'd

- Asymptotic normality $\Rightarrow$ we can use standard methods of inference for $\boldsymbol{\beta}^*$ as in the regression case, e.g., Wald-type intervals etc.

- Unlike regression case, measures of performance are not well-behaved

- Recall our measures of performance

  - Population-level error: $R \triangleq P1_{Y\mathbf{X}^\top \boldsymbol{\beta}^* < 0}$

  - Conditional error: $C(\widehat{\boldsymbol{\beta}}_n) \triangleq P1_{Y\mathbf{X}^\top \widehat{\boldsymbol{\beta}}_n < 0}$

  - Average error: $A_n \triangleq \mathbb{E}C(\widehat{\boldsymbol{\beta}}_n)$

Comparing measures of performance

▶ Fact: the three measures of performance need not coincide even in infinite samples

Blank page for notes

Blank page for notes

Thank you.

`eric.laber@duke.edu`

`laber-labs.com`