Logistic Regression (Reading: 17.1 - 17.5)

Learning goals:

- Understand similarities and differences between classification and regression.
- Introduce the logistic model and the cross-entropy loss.

UC Berkeley Data 100 Summer 2019
Sam Lau

(Slides adapted from John DeNero)

Announcements

- HW6 due **Tuesday**
- Project 2 out **Tuesday**
  - Due the following Tuesday, Aug 5.

Classification

Classification

Classifiers are functions used to make predictions about a **categorical** variable.

- *Kawhi Leonard takes a 21-foot jump shot against the 76ers. Will it go in?*
- *A 21-year-old white female from Florida is arrested for assault a second time.*
  If let free, will she be arrested again for a violent crime in the next two years?

Classification and Regression

Classification is not so different from regression:

- Fit a model using labeled training examples (x, y), then apply it to unlabeled examples x.
- We assume unlabeled examples have similar labels.

And, we usually have the following questions:

- Is association between (x, y) in training set representative?
- Are there enough training examples?
- Will the model generalize?

Linear Regression Review

Prediction problem: Predict y from covariates (features) x.

Regression: Estimate $f^*(x)$ for unknown distribution over (X, Y).

Linear Regression: Assume $f^*(x) = \theta^* \cdot x$ and estimate θ^*, a vector of parameters.

Model: Set of all distributions $\theta \cdot x$ you can get by choosing θ.

To Fit Model: Choose a loss, then minimize loss.

Linear Regression Review

Squared Loss for Linear Regression: $(y - \theta \cdot x)^2$

Average Loss / Empirical Risk:

For training (i.e. learning) set of observations $(\boldsymbol{x_1}, y_1), \ldots, (\boldsymbol{x_n}, y_n)$

$$L(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{y}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \boldsymbol{\theta} \cdot \boldsymbol{X_i})^2 + \lambda \sum_{j=1}^{p} |\theta_j|$$

Regularization: Add a term to average loss that encourages small θ.

Classification

Classification prediction problem: Predict y from features x.

- Now, y in a fixed set of possible classes, e.g., {*make*, *miss*}.
- Suppose we assign *make* = 1 and *miss* = 0.
    - Two classes = **binary classification problem**
- Intuition: y feels like a Bernoulli RV with p dependent on x!
- Let's use (X, Y) to refer to the pair of RVs drawn from population. X contains features, Y contains the true value.
- Interested in **E(Y | X)**: if I know X, what is Y (on average)?

Classification

$$E(Y|X) = 1 \cdot P(Y = 1|X) + 0 \cdot P(Y = 0|X)$$
$$= P(Y = 1|X) = f^*_{\boldsymbol{\theta}}(X)$$

- Want to estimate P(Y = 1 | X). This is numeric!
- Intuition: Tweak regression to predict probabilities.
- Linear Regression: f(x) = θ·x
- Logistic Regression: $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sigma(\boldsymbol{\theta} \cdot \boldsymbol{x})$

Linear Regression

With a tweak

(Demo)

Logistic Regression

[bit.ly/at-d100](bit.ly/at-d100)

Logistic Regression for Binary Classification

- As usual, we pick model + loss, then fit with GD.

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sigma(\boldsymbol{\theta} \cdot \boldsymbol{x}) \qquad \text{where } \sigma(t) = \frac{1}{1 + \exp(-t)}$$

Let $z_i = f_{\boldsymbol{\theta}}(\boldsymbol{X_i})$.

$$L(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{y}) = -\frac{1}{n} \sum_{i=1}^{n} [y_i \log z_i + (1 - y_i) \log(1 - z_i)] + \sum_{j=1}^{p} \theta_j^2$$

- To predict category, set a **decision rule**:
    - E.g. if f(x) ≥ 0.5, predict 1

(Demo)

Why use the sigmoid function?

- The function σ(t) is called the **sigmoid** (or **logistic**) function.
- Arises from assuming that the log-odds ratio is linear:

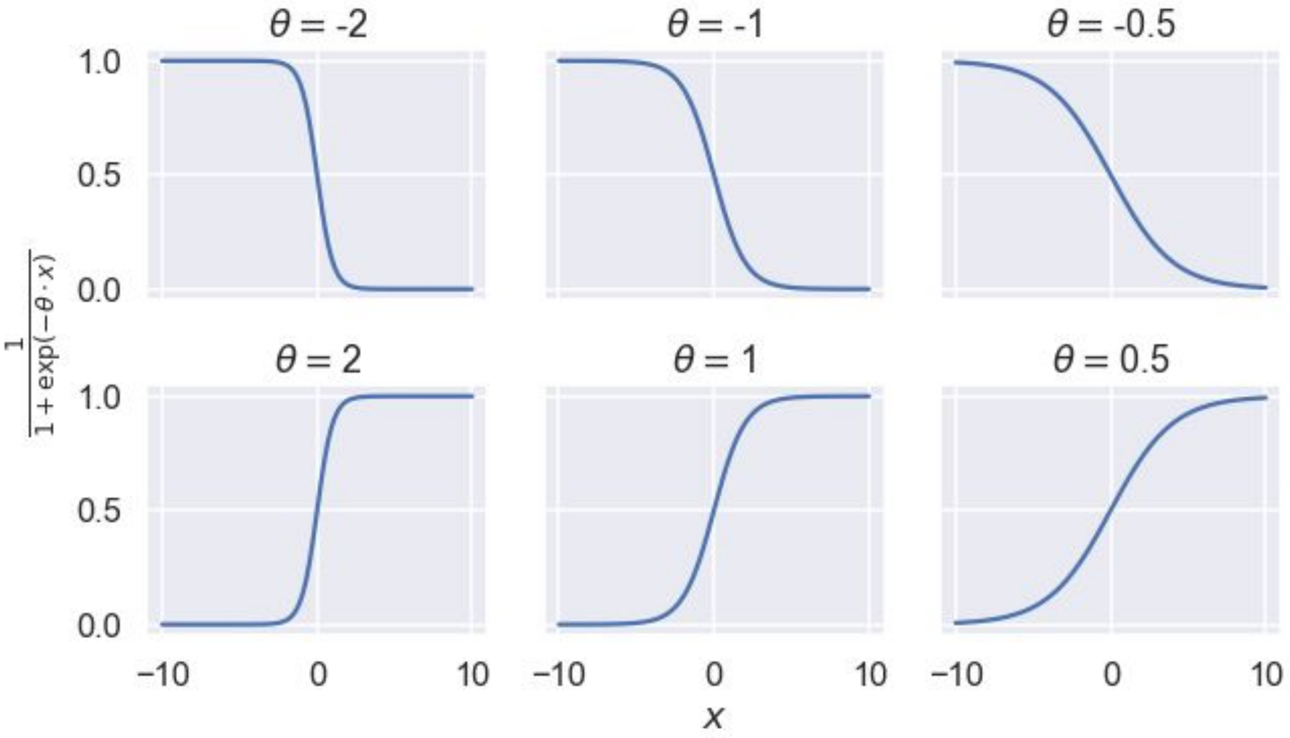
$$\log\left(\frac{P(Y=1|X)}{P(Y=0|X)}\right) = X \cdot \theta \quad \Longrightarrow \quad P(Y=1|X) = \frac{1}{1+\exp(-X \cdot \theta)}$$

$$= \sigma(X \cdot \theta)$$

Why use the sigmoid function?

θ feels like the "slope" of logistic model.

If θ is +, higher values of x give higher probabilities.



Why Not Squared Loss?

$$L(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{y}) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log z_i + (1 - y_i)\log(1 - z_i)]$$

- We use average **cross-entropy loss** for logistic regression.
- Squared loss actually not a terrible choice.
- However, the corresponding empirical risk function can be non-convex, and therefore difficult to minimize.
- Cross-entropy loss has other motivations that you can learn about in a machine learning or probability course: maximum likelihood or minimum KL-divergence.

Practice with Cross Entropy Loss

$$L(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{y}) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log z_i + (1 - y_i)\log(1 - z_i)]$$

$$\ell(\boldsymbol{\theta}, \boldsymbol{x}, y_i) = -y_i \log z_i - (1 - y_i)\log(1 - z_i)$$

Suppose we only have one feature (and no intercept).

If θ = 2, what is the point loss for:

- (x, y) = (5, 1)?
- (0, 1)? (0, 0)?
- (-1, 1)?
- If running SGD, in what direction would θ be updated?

Practice with Cross Entropy Loss

$$\ell(\boldsymbol{\theta}, \boldsymbol{x}, y_i) = -y_i \log z_i - (1 - y_i) \log(1 - z_i)$$

$$\theta = 2$$

For $(5, 1) : z = \sigma(5 \cdot 2) \approx 1$ $\quad$ $\ell = -1 \cdot \log 1 - 0 = 0$

For $(0, 1) : z = \sigma(0 \cdot 2) = \frac{1}{2}$ $\quad$ $\ell = -1 \cdot \log \frac{1}{2} - 0 = 0.693$

For $(1, 1) : z = \sigma(0 \cdot 2) = \frac{1}{2}$ $\quad$ $\ell = 0 - 1 \cdot \log \frac{1}{2} = 0.693$

For $(-1, 1) : z = \sigma(-1 \cdot 2) = 0.119$ $\quad$ $\ell = 0 - 1 \cdot \log 0.119 = 2.129$

Notice how one term in the loss is always 0?

Demo: Logistic Regression


NBA
SIMMONS
25
21
2
ALL-ANGLES

Summary

- Classification can be framed as a regression problem.
- Logistic regression uses two new pieces of machinery:
  - The logistic model: arises from assumption on probabilities.
  - The cross-entropy loss: convex for the logistic model.

The wrong predictions can be the most interesting ones!