Logistic regression

Daniel Hsu (COMS 4771)

The logistic regression model

Logistic regression is a model for binary classification data with feature vectors in $\mathbb{R}^d$ and labels in $\{-1, +1\}$. Data $(\boldsymbol{X}_1, Y_1), \dots, (\boldsymbol{X}_n, Y_n)$ are treated as iid random variables taking values in $\mathbb{R}^d \times \{-1, +1\}$, and for each $\boldsymbol{x} \in \mathbb{R}^d$,

$$Y_i \mid \boldsymbol{X}_i = \boldsymbol{x} \sim \operatorname{Bern}(\sigma(\boldsymbol{x}^\mathsf{T} \boldsymbol{w}))$$

where $\sigma(t) = 1/(1 + \exp(-t))$ is the sigmoid function. Here, $\boldsymbol{w} \in \mathbb{R}^d$ is the parameter of the model, and it is not involved in the marginal distribution of $\boldsymbol{X}_i$ (which we leave unspecified).

Maximum likelihood

The log-likelihood of $\boldsymbol{w}$ given $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_n, y_n) \in \mathbb{R}^d \times \{-1, +1\}$ is

$$-\sum_{i=1}^n \ln(1 + \exp(-y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{w})) + (\text{terms that do not involve } \boldsymbol{w}).$$

There is no closed-form expression for the maximizer of the log-likelihood. Nevertheless, we can approximately minimize the negative log-likelihood with gradient descent.

Empirical risk minimization

Maximum likelihood is very different from finding the linear classifier of smallest empirical zero-one loss risk. Finding the empirical zero-one loss risk minimizer is computationally intractable in general.

Finding a linear separator

There are special cases when finding the empirical zero-one loss risk minimizer is computationally tractable. One is when the training data is *linearly separable*: i.e., when there exists $\boldsymbol{w}^\star \in \mathbb{R}^d$ such that

$$y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{w}^\star > 0, \quad \text{for all } i = 1, \dots, n.$$

Claim. Define $L(\boldsymbol{w}) := \sum_{i=1}^n \ln(1 + \exp(-y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{w}))$. Suppose $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_n, y_n) \in \mathbb{R}^d \times \{-1, +1\}$ is linearly separable. Then any $\hat{\boldsymbol{w}} \in \mathbb{R}^d$ with

$$L(\hat{\boldsymbol{w}}) < \inf_{\boldsymbol{w} \in \mathbb{R}^d} L(\boldsymbol{w}) + \ln(2)$$

is a linear separator.

Proof. We first observe that the infimum[1] (i.e., greatest lower bound) of L is zero. Let $\boldsymbol{w}^\star \in \mathbb{R}^d$ be a linear separator, so $s_i := y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{w}^\star > 0$ for all $i = 1, \dots, n$. For any $r > 0$,

$$L(r\boldsymbol{w}^\star) = \sum_{i=1}^n \ln(1 + \exp(-r s_i)),$$

[1] https://en.wikipedia.org/wiki/Infimum_and_supremum

and therefore

$$\lim_{r \to \infty} \sum_{i=1}^{n} \ln(1 + \exp(-r s_i)) = 0.$$

Every term $\ln(1 + \exp(-y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{w}))$ in $L(\boldsymbol{w})$ is positive, so $L(\boldsymbol{w}) > 0$. Therefore, we conclude that

$$\inf_{\boldsymbol{w} \in \mathbb{R}^d} L(\boldsymbol{w}) = 0.$$

So now we just have to show that any $\hat{\boldsymbol{w}} \in \mathbb{R}^d$ with

$$L(\hat{\boldsymbol{w}}) < \ln(2)$$

is a linear separator. So let $\hat{\boldsymbol{w}}$ satisfy $L(\hat{\boldsymbol{w}}) < \ln(2)$, which implies

$$\ln(1 + \exp(-y_i \boldsymbol{x}_i^\mathsf{T} \hat{\boldsymbol{w}})) < \ln(2)$$

for every $i = 1, \ldots, n$. Exponentiating both sides gives

$$1 + \exp(-y_i \boldsymbol{x}_i^\mathsf{T} \hat{\boldsymbol{w}}) < 2.$$

Now subtracting 1 from both sides and taking logarithms gives

$$-y_i \boldsymbol{x}_i^\mathsf{T} \hat{\boldsymbol{w}} < 0.$$

This means that $\hat{\boldsymbol{w}}$ correctly classifies $(\boldsymbol{x}_i, y_i)$. Since this holds for all $i = 1, \ldots, n$, it follows that $\hat{\boldsymbol{w}}$ is a linear separator. $\square$

Surrogate loss

Even if $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n) \in \mathbb{R}^d \times \{-1, +1\}$ is not linearly separable, approximately maximizing the log-likelihood can yield a good linear classifier. This is because maximizing L is the same as minimizing the *empirical logistic loss risk*:

$$\widehat{\mathcal{R}}(\boldsymbol{w}) := \frac{1}{n} \sum_{i=1}^{n} \ell_{\log}(y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{w})$$

where

$$\ell_{\log}(z) := -\ln \sigma(z)$$

is the *logistic loss*. The logistic loss (up to scaling) turns out to be an upper-bound on the zero-one loss:

$$\ell_{\text{zo}}(z) \leq \frac{1}{\ln 2} \ell_{\log}(z),$$

where $\ell_{\text{zo}}(z) = \mathbb{1}_{\{z \leq 0\}}$. If the empirical logistic loss risk is small, then the empirical zero-one loss is also small.

Gradient descent for logistic regression

The derivative of $\ell_{\log}$ is given by

$$\begin{aligned} \frac{\mathrm{d}\ell_{\log}(z)}{\mathrm{d}z} &= -\frac{1}{\sigma(z)} \cdot \frac{\mathrm{d}\sigma(z)}{\mathrm{d}z} \\ &= -\frac{1}{\sigma(z)} \cdot \sigma(z) \cdot \sigma(-z) \\ &= -\sigma(-z). \end{aligned}$$

Therefore, by linearity and the chain rule, the negative gradient of $\widehat{\mathcal{R}}$ with respect to $\boldsymbol{w}$ is

$$\begin{aligned} -\nabla \widehat{\mathcal{R}}(\boldsymbol{w}) &= -\frac{1}{n}\sum_{i=1}^{n} \nabla\, \ell_{\log}(y_i \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{w}) \\ &= -\frac{1}{n}\sum_{i=1}^{n} \left.\frac{\mathrm{d}\ell_{\log}(z)}{\mathrm{d}z}\right|_{z=y_i\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{w}} \cdot \nabla \left(y_i \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{w}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \sigma(-y_i \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{w}) \cdot y_i \boldsymbol{x}_i. \end{aligned}$$

Now suppose $\boldsymbol{A} = [\boldsymbol{x}_1 | \cdots | \boldsymbol{x}_n]^{\mathsf{T}} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{b} = [y_1 | \cdots | y_n]^{\mathsf{T}} \in \mathbb{R}^n$. (Notice that we have omitted the $1/\sqrt{n}$ scaling that we had for least squares linear regression.) Then the negative gradient of $\widehat{\mathcal{R}}$ can be written as

$$-\nabla \widehat{\mathcal{R}}(\boldsymbol{w}) = \frac{1}{n}\boldsymbol{A}^{\mathsf{T}}(\boldsymbol{b} \odot \sigma(-\boldsymbol{b} \odot (\boldsymbol{A}\boldsymbol{w}))),$$

where $\boldsymbol{u} \odot \boldsymbol{v} \in \mathbb{R}^n$ is the coordinate-wise product of vectors $\boldsymbol{u} \in \mathbb{R}^n$ and $\boldsymbol{v} \in \mathbb{R}^n$, and $\sigma(\boldsymbol{v}) \in \mathbb{R}^n$ is the coordinate-wise application of the sigmoid function to $\boldsymbol{v} \in \mathbb{R}^n$.

Gradient descent for logistic regression begins with an initial weight vector $\boldsymbol{w}^{(0)} \in \mathbb{R}^d$, and then iteratively updates it by subtracting a positive multiple $\eta > 0$ of the gradient at the current iterate:

$$\boldsymbol{w}^{(t)} := \boldsymbol{w}^{(t-1)} - \eta \nabla \widehat{\mathcal{R}}(\boldsymbol{w}^{(t-1)}).$$