Machine Learning

Regression

Mayson Ma

Courant Institute of Mathematical Sciences - New York University

June 2021

Table of Contents

Linear Regression

Symbols

Sample points $x_1, x_2, ...x_n \in \mathbb{R}$ and $X \in \mathbb{R}^{n\times d}$ is the design matrix of sample points. The associated labels are $y_1, y_2, ..., y_n \in \mathbb{R}$, and $y = [y_1 \quad \cdots \quad y_n]^\top \in \mathbb{R}^n$.

Hypothesis set

Linear functions

$$\left\{x \mapsto w^\top x + \alpha : x \in \mathbb{R}^d, \alpha \in \mathbb{R}\right\}$$

Optimization Problems

Empirical risk minimization

$$\min_{w,\alpha} F(w, \alpha) = ||(Xw + \alpha) - y||^2$$

Solution

Use the fictitious dimension trick, rewrite the objective functions:

$$\min_w F(w) = ||Xw - y||^2$$

$$\text{where } X = \begin{bmatrix} x_1^\top & 1 \\ \vdots & \vdots \\ x_n^\top & 1 \end{bmatrix} \in \mathbb{R}^{n\times(d+1)},\ w = \begin{bmatrix} w_1 \\ \vdots \\ w_d \\ \alpha \end{bmatrix} \in \mathbb{R}^{d+1}$$

Solve by calculus

The objective is a convex and differentiable function

$$F(w) = w^\top X^\top Xw - 2y^\top Xw + y^\top y$$

$$\Delta_w F = 2X^\top Xw - 2X^\top y$$

Set $\Delta_w F = 0 \Leftrightarrow X^T Xw = X^T y$

Solution

The normal equation

$$X^T X w = X^T y$$

Solution to the normal equation

If $X^\top X$ is not singular, $w = (X^T X)^{-1} X^T y$.

If $X^\top X$ is singular, the problem is underconstrained, $w = X^+ y$ in general, where X^+ is the pseudoinverse of X.

The Least Norm Solution

Proposition

The solution $w^+ = X^+ y$ has two properties:

- $w = w^+$ is a minimizer of $|Xw - y|^2$
- If another $\widehat{w}$ achieves that minimum, then $|w^+| < |\widehat{w}|$

Let $X = U \Sigma V^\intercal$ be the SVD decomposition of X. Then,

$$w^+ = X^+ y = V \Sigma^+ U^\intercal y = \sum_{i:\sigma_i > 0} \frac{1}{\sigma_i} v_i (u_i^\intercal y)$$

Lemma

w^+ is in the row space of X.

Proof. From the SVD, we will have r positive singular values in descending order $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r > 0$. The corresponding v's forms the basis for the row space of X. The last $n - r$ v's are in the nullspace of X. **QED.**

The Least Norm Solution

Now we that $w^+ = X^+y$ **is the minimum norm least squares solution**. When X has independent columns and rand $r = n$, this is the only least squares solution. But if there are nonzero vectors w' in the nullspace of X, they can be added to w^+. The error $|y - X(w^+ + w')|$ is not affected since $Xw' = 0$. But the norm $|w^+ + w'|^2$ will grow to $|w^+|^2 + |w'|^2$. Those pieces are orthogonal: Row space $\perp$ nullspace.

Ridge Regression

Motivation

If the design matrix $X \in \mathbb{R}^{n \times d}$ has dependent columns and $Xw = 0$ has nonzero solutions, then $X^\intercal X$ cannot be invertible. This is where we need X^+. A gentle approach will **regularize** least squares by adding penalty term.

Optimization problem

Minimize $|Xw - y|^2 + \lambda|w|^2$
Solve $(X^\intercal X + \lambda I)w = X^\intercal y$

Controlling the complexity of $\widehat{w}$

- Increasing the ridge parameter λ shrinks the norm $|\widehat{w}|$.
- Even as $\lambda \to 0$, $\widehat{w}$ picks out the least norm least squares solution.

The Pseudoinverse X^+ is the Limit of $(X^\intercal X + \lambda I)^{-1} X^\intercal$

The solution to the ridge regression objective function is

$$\widehat{w} = (X^\intercal X + \lambda I)^{-1} X^\intercal y$$

Let $X = U\Sigma V^\intercal$ be the SVD decomposition of X. Then,

$$X^\intercal X + \lambda I = V\Sigma^2 V^\intercal + \lambda I = V(\Sigma^2 + \lambda I)V^\intercal$$

since V is square orthogonal matrix ($V^\intercal = V^{-1}$)

$$\widehat{w} = \mathrm{V}[(\Sigma^2 + \lambda I)^{-1}\Sigma]\mathrm{U}^\intercal y = \sum_{i=1}^{d} \frac{\sigma_i}{\sigma_i^2 + \lambda} v_i(u_i^\intercal y)$$

$$\lim_{\lambda \to 0} \widehat{w} = \sum_{i=1}^{d} \lim_{\lambda \to 0} \frac{\sigma_i}{\sigma_i^2 + \lambda} v_i(u_i^\intercal y) = \sum_{i:\sigma_i > 0} \frac{1}{\sigma_i} v_i(u_i^\intercal y) = w^+$$

Logistic Regression

Symbols

Sample points $x_1, x_2, ...x_n \in \mathbb{R}$ and $X \in \mathbb{R}^{n \times d}$ is the design matrix of sample points. The associated labels are $y_1, y_2, ..., y_n \in \{0, 1\}$, and $y = [y_1 \quad \cdots \quad y_n]^\top \in \{0, 1\}^n$.

Model

Let $\sigma(z) = \frac{1}{1+e^{-z}}$ be the sigmoid function. Then,

$$p_1 = \Pr(y = 1|x) = \sigma(w^\top x) = \frac{1}{1 + e^{-w^\top x}}$$

$$p_0 = \Pr(y = 0|x) = 1 - \sigma(w^\top x) = \frac{e^{-w^\top x}}{1 + e^{-w^\top x}}$$

Combine p_0 and p_1 gives $\Pr(y|x) = p_1^y \cdot p_0^{(1-y)}$.

Logistic Regression

MLE

$$
\begin{aligned}
\hat{w} &= \underset{w}{\operatorname{argmax}} \ln \prod_{i=1}^{n} \Pr(y_i | x_i) \\
&= \underset{w}{\operatorname{argmax}} \ln \prod_{i=1}^{n} \Pr(y_i = 1 | x_i)^{y_i} \cdot \Pr(y_i = 0 | x_i)^{(1-y_i)} \\
&= \underset{w}{\operatorname{argmax}} \sum_{i=1}^{n} y_i \ln \sigma(w^\top x_i) + (1 - y_i) \ln(1 - \sigma(w^\top x_i))
\end{aligned}
$$

Optimization Problem

$\min_w F(w) = -\sum_{i=1}^{n} y_i \ln \sigma(w^\top x_i) + (1 - y_i) \ln(1 - \sigma(w^\top x_i))$

Solution

Note that $\sigma'(z) = \sigma(z)(1 - \sigma(z))$.
Let $z_i = \sigma(w^\top x_i)$

$$\begin{aligned}\Delta_w F &= -\sum_{i=1}^{n} \left(\frac{y_i}{z_i} - \frac{1 - y_i}{1 - z_i} \right) z_i(1 - z_i)x_i \\ &= -\sum_{i=1}^{n} (y_i - z_i)x_i \\ &= -X^\top(y - \sigma(Xw))\end{aligned}$$

Solve by gradient descent

$w \leftarrow w + \epsilon \cdot X^\top(y - \sigma(Xw))$, where ϵ is the learning rate.

Newton's Method

Idea

You are at point v. Approximate $F(w)$ near v by a quadratic function. Jump to its unique critical point. Repeat until bored.

Math

Taylor series at v:

$$F(v+d) \approx F(v) + \Delta F(v)^\top d + \frac{1}{2} d^\top \Delta^2 F(v) d$$

where $\Delta^2 F(v)$ is the **Hessian matrix** of F at point v.

Take derivative w.r.t d, set it to 0 and solve for d, i.e. find the critical point:

$$\Delta F(v) + \Delta^2 F(v) d = 0 \Rightarrow d = -(\Delta^2 F(v))^{-1} \Delta F(v)$$

Newton's Method

Algorithm

pick starting point w

repeat

$e \leftarrow$ solution to linear system $(\Delta^2 F(w))^{-1} e = -\Delta F(w)$

$w \leftarrow w + e$

until convergence

Comments

- Iterative optimization method for smooth $F(w)$
- Often much faster than gradient descent
- Does not know the difference between minima, maxima or saddle points
- Starting point must be "close enough" to desired solution

Solution

Solve by Newton's method

- Recall that $\Delta_w F = X^\intercal(z - y)$, where $z \in \mathbb{R}^n$ and $z_i = \sigma(w^\intercal x_i)$
- So $\Delta_W^2 F$ resolves to $X^\intercal(\Delta_w z)$
- Note that the i-th row of $\Delta_w z$ will be $z_i(1 - z_i)x_i^\intercal$. So we can rewrite this term as

$$\Delta_w z = \Omega X$$

where $\Omega = \begin{bmatrix} z_1(1-z_1) & & & \\ & z_2(1-z_2) & & \\ & & \ddots & \\ & & & z_n(1-z_n) \end{bmatrix}$

- Finally, $\Delta_W^2 F = X^\intercal \Omega X$

Solve by Newton's method

Algorithm

$w \leftarrow 0$
repeat
  $e \leftarrow$ solution to normal equations $(X^\top \Omega X)e = X^\top(y - z)$
  $w \leftarrow w + e$
until convergence

An example of iteratively reweighted least squares

- Ω is positive definite $\Rightarrow X^\top \Omega X$ is positive semidefinite $\Rightarrow F(w)$ is convex
- Ω prioritizes points with z near 0.5; tunes out points near 0 or 1.
- If n very large, save time by using a random subsample of the points per iteration. Increase sample size as you go.