

Linear and logistic regression



École des Ponts
ParisTech

Guillaume Obozinski

Ecole des Ponts - ParisTech



Master MVA

Outline

- 1 Linear regression
- 2 Logistic regression
- 3 Fisher discriminant analysis

Linear regression

Design matrix

Consider a finite collection of vectors $x_i \in \mathbb{R}^d$ pour $i = 1 \dots n$.

Design Matrix

$$X = \begin{bmatrix} \text{---} & x_1^\top & \text{---} \\ \vdots & \vdots & \vdots \\ \text{---} & x_n^\top & \text{---} \end{bmatrix}.$$

We will assume that the vectors are *centered*, i.e. that $\sum_{i=1}^n x_i = 0$, and *normalized*, i.e. that $\frac{1}{n} \sum_{i=1}^n x_i^2 = 1$.

If x_i are not centered the design matrix of centered data can be constructed with the rows $x_i - \bar{x}^\top$ with $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.

Normalization usually consists in dividing each row by its empirical standard deviation.

Generative models vs conditional models

- X is the input variable
- Y is the output variable

A **generative model** is a model of the joint distribution $p(x, y)$.

A **conditional model** is a model of the conditional distribution $p(y|x)$.

Conditional models *vs* Generative models

- CM make fewer assumptions about the data distribution
- CM require fewer parameters
- CM are typically computationally harder to learn
- CM can typically not handle missing data or latent variables

Probabilistic version of linear regression

Modeling the conditional distribution of Y given X by

$$Y | X \sim \mathcal{N}(\mathbf{w}^\top X + b, \sigma^2)$$

or equivalently $Y = \mathbf{w}^\top X + b + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

The offset can be ignored up to a reparameterization.

$$Y = \tilde{\mathbf{w}}^\top \begin{pmatrix} x \\ 1 \end{pmatrix} + \epsilon.$$

Likelihood for one pair

$$p(y_i | \mathbf{x}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{\sigma^2}\right)$$

Negative log-likelihood

$$-\ell(\mathbf{w}, \sigma^2) = -\sum_{i=1}^n \log p(y_i | \mathbf{x}_i) = \frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2} \sum_{i=1}^n \frac{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{\sigma^2}.$$

Probabilistic version of linear regression

$$\min_{\sigma^2, \mathbf{w}} \frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2} \sum_{i=1}^n \frac{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{\sigma^2}$$

The minimization problem in \mathbf{w}

$$\min_{\mathbf{w}} \frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$$

that we recognize as the usual linear regression, with

- $\mathbf{y} = (y_1, \dots, y_n)^\top$ and
- \mathbf{X} the design matrix with rows equal to \mathbf{x}_i^\top .

Optimizing over σ^2 , we find:

$$\hat{\sigma}_{MLE}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\mathbf{w}}_{MLE}^\top \mathbf{x}_i)^2$$

Solving linear regression

To solve $\min_{\mathbf{w} \in \mathbb{R}^p} \widehat{\mathcal{R}}_n(f_{\mathbf{w}})$, we consider that

$$\widehat{\mathcal{R}}_n(f_{\mathbf{w}}) = \frac{1}{2n} (\mathbf{w}^\top \mathbf{X}^\top \mathbf{X} \mathbf{w} - 2 \mathbf{w}^\top \mathbf{X}^\top \mathbf{y} + \|\mathbf{y}\|^2)$$

is a **differentiable convex** function whose minima are thus characterized by the

Normal equations

$$\mathbf{X}^\top \mathbf{X} \mathbf{w} - \mathbf{X}^\top \mathbf{y} = \mathbf{0}$$

If $\mathbf{X}^\top \mathbf{X}$ is invertible, then $\widehat{\mathbf{w}}_{MLE}$ is given by:

$$\widehat{\mathbf{w}}_{MLE} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

Problem: $\mathbf{X}^\top \mathbf{X}$ is never invertible for $p > n$ and thus the solution is not unique (and any solution is overfit).

Logistic regression

Logistic regression

Classification setting:

$$\mathcal{X} = \mathbb{R}^p, \mathcal{Y} \in \{0, 1\}.$$

Key assumption:

$$\log \frac{\mathbb{P}(Y = 1 | X = \mathbf{x})}{\mathbb{P}(Y = 0 | X = \mathbf{x})} = \mathbf{w}^\top \mathbf{x}$$

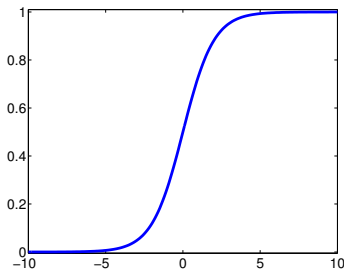
Implies that

$$\mathbb{P}(Y = 1 | X = \mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x})$$

for

$$\sigma : z \mapsto \frac{1}{1 + e^{-z}},$$

the **logistic function**.



- The logistic function is part of the family of *sigmoid functions*.
- Often called “the” sigmoid function.

Properties:

$$\begin{aligned} \forall z \in \mathbb{R}, \quad \sigma(-z) &= 1 - \sigma(z), \\ \forall z \in \mathbb{R}, \quad \sigma'(z) &= \sigma(z)(1 - \sigma(z)) \\ &= \sigma(z)\sigma(-z). \end{aligned}$$

Likelihood for logistic regression

Let $\eta := \sigma(\mathbf{w}^\top \mathbf{x} + b)$. W.l.o.g. we assume $b = 0$.

By assumption: $Y|X = \mathbf{x} \sim \text{Ber}(\eta)$.

Likelihood

$$p(Y = y|X = \mathbf{x}) = \eta^y(1 - \eta)^{1-y} = \sigma(\mathbf{w}^\top \mathbf{x})^y \sigma(-\mathbf{w}^\top \mathbf{x})^{1-y}.$$

Log-likelihood

$$\begin{aligned}\ell(\mathbf{w}) &= y \log \sigma(\mathbf{w}^\top \mathbf{x}) + (1 - y) \log \sigma(-\mathbf{w}^\top \mathbf{x}) \\ &= y \log \eta + (1 - y) \log(1 - \eta) \\ &= y \log \frac{\eta}{1 - \eta} + \log(1 - \eta) \\ &= y \mathbf{w}^\top \mathbf{x} + \log \sigma(-\mathbf{w}^\top \mathbf{x})\end{aligned}$$

Maximizing the log-likelihood

Log-likelihood of a sample

Given an i.i.d. training set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$

$$\ell(\mathbf{w}) = \sum_{i=1}^n y_i \mathbf{w}^\top \mathbf{x}_i + \log \sigma(-\mathbf{w}^\top \mathbf{x}_i).$$

The log-likelihood is differentiable and concave.

\Rightarrow Its global maxima are its stationary points.

Gradient of ℓ

$$\begin{aligned} \nabla \ell(\mathbf{w}) &= \sum_{i=1}^n y_i \mathbf{x}_i - \mathbf{x}_i \frac{\sigma(-\mathbf{w}^\top \mathbf{x}_i) \sigma(\mathbf{w}^\top \mathbf{x}_i)}{\sigma(-\mathbf{w}^\top \mathbf{x}_i)} \\ &= \sum_{i=1}^n (y_i - \eta_i) \mathbf{x}_i \quad \text{with} \quad \eta_i = \sigma(\mathbf{w}^\top \mathbf{x}_i). \end{aligned}$$

Thus, $\nabla \ell(\mathbf{w}) = 0 \Leftrightarrow \sum_{i=1}^n \mathbf{x}_i (y_i - \sigma(\mathbf{w}^\top \mathbf{x}_i)) = 0.$

No closed form solution !

Second order Taylor expansion

Need an iterative method to solve
$$\sum_{i=1}^n \mathbf{x}_i (y_i - \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)) = 0.$$

→ Gradient descent (aka steepest descent)

→ Newton's method

Hessian of ℓ

$$\begin{aligned} H\ell(\mathbf{w}) &= \sum_{i=1}^n \mathbf{x}_i (0 - \sigma'(\mathbf{w}^\top \mathbf{x}_i)) \sigma'(-\mathbf{w}^\top \mathbf{x}_i) \mathbf{x}_i^\top \\ &= \sum_{i=1}^n -\eta_i (1 - \eta_i) \mathbf{x}_i \mathbf{x}_i^\top = -\mathbf{X}^\top \text{Diag}(\eta_i (1 - \eta_i)) \mathbf{X} \end{aligned}$$

where \mathbf{X} is the design matrix.

→ Note that $-H\ell$ is p.s.d. $\Rightarrow \ell$ is concave.

Newton's method

Use the Taylor expansion

$$\ell(\mathbf{w}^t) + (\mathbf{w} - \mathbf{w}^t)^\top \nabla \ell(\mathbf{w}^t) + \frac{1}{2} (\mathbf{w} - \mathbf{w}^t)^\top H \ell(\mathbf{w}^t) (\mathbf{w} - \mathbf{w}^t).$$

and minimize w.r.t. \mathbf{w} . Setting $\mathbf{h} = \mathbf{w} - \mathbf{w}^t$, we get

$$\max_{\mathbf{h}} \mathbf{h}^\top \nabla_{\mathbf{w}} \ell(\mathbf{w}) + \frac{1}{2} \mathbf{h}^\top H \ell(\mathbf{w}) \mathbf{h}.$$

I.e., for logistic regression, writing $\mathbf{D}_\eta = \text{Diag}((\eta_i(1 - \eta_i)))_i$

$$\min_{\mathbf{h}} \mathbf{h}^\top \mathbf{X}^\top (\mathbf{y} - \boldsymbol{\eta}) - \frac{1}{2} \mathbf{h}^\top \mathbf{X}^\top \mathbf{D}_\eta \mathbf{X} \mathbf{h}$$

Modified normal equations

$$\mathbf{X}^\top \mathbf{D}_\eta \mathbf{X} \mathbf{h} - \mathbf{X}^\top \tilde{\mathbf{y}} \quad \text{with} \quad \tilde{\mathbf{y}} = \mathbf{y} - \boldsymbol{\eta}.$$

Iterative Reweighted Least Squares (IRLS)

Assuming $\mathbf{X}^\top \mathbf{D}_\eta \mathbf{X}$ is invertible, the algorithm takes the form

$$\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} + (\mathbf{X}^\top \mathbf{D}_{\eta^{(t)}} \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{y} - \boldsymbol{\eta}^{(t)}).$$

This is called iterative reweighted least squares because each step is equivalent to solving the reweighted least squares problem:

$$\frac{1}{2} \sum_{i=1}^n \frac{1}{\tau_i^2} (\mathbf{x}_i^\top \mathbf{h} - \check{y}_i)^2$$

with

$$\tau_i^2 = \frac{1}{\eta_i^{(t)} (1 - \eta_i^{(t)})} \quad \text{and} \quad \check{y}_i = \tau_i^2 (y_i - \eta_i^{(t)}).$$

Alternate formulation of logistic regression

If $y \in \{-1, 1\}$, then

$$\mathbb{P}(Y = y|X = \mathbf{x}) = \sigma(y \mathbf{w}^\top \mathbf{x})$$

Log-likelihood

$$\ell(\mathbf{w}) = \log \sigma(y \mathbf{w}^\top \mathbf{x}) = -\log(1 + \exp(-y \mathbf{w}^\top \mathbf{x}))$$

Log-likelihood for a training set

$$\ell(\mathbf{w}) = -\sum_{i=1}^n \log(1 + \exp(-y_i \mathbf{w}^\top x_i))$$

Fisher discriminant analysis

Generative classification

$X \in \mathbb{R}^p$ and $Y \in \{0, 1\}$. Instead of modeling directly $p(y | \mathbf{x})$ model $p(y)$ and $p(\mathbf{x} | y)$ and deduce $p(y | \mathbf{x})$ using Bayes rule.

In classification $\mathbb{P}(Y = 1 | X = \mathbf{x}) =$

$$\frac{\mathbb{P}(X = \mathbf{x} | Y = 1) \mathbb{P}(Y = 1)}{\mathbb{P}(X = \mathbf{x} | Y = 1) \mathbb{P}(Y = 1) + \mathbb{P}(X = \mathbf{x} | Y = 0) \mathbb{P}(Y = 0)}$$

For example one can assume

- $\mathbb{P}(Y = 1) = \pi$
- $\mathbb{P}(X = \mathbf{x} | Y = 1) \sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$
- $\mathbb{P}(X = \mathbf{x} | Y = 0) \sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$.

Fisher's discriminant aka Linear Discriminant Analysis (LDA)

Previous model with the constraint $\Sigma_1 = \Sigma_0 = \Sigma$. Given a training set, the different model parameters can be estimated using the maximum likelihood principle, which leads to

$$(\hat{\pi}, \hat{\mu}_1, \hat{\mu}_0, \hat{\Sigma}_1, \hat{\Sigma}_0).$$