Introduction to Machine learning

Sylvain Le Corff

Logistic regression

Summary

Supervised Learning

Supervised Learning Framework

- Input measurement $\mathbf{X} \in \mathcal{X}$ (often $\mathcal{X} \subset \mathbb{R}^d$).
- Output measurement $Y \in \mathcal{Y}$.
- The joint distribution of $(\mathbf{X}, Y)$ is unknown.
- $Y \in \{1, \ldots, M\}$ (classification) or $Y \in \mathbb{R}^m$ (regression).
- A predictor is a measurable function $f : \mathcal{X} \to \mathcal{Y}$.

Training data

- i.i.d. with the same distribution as $(\mathbf{X}, Y)$:

$$\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}.$$

Goal

- Construct a good predictor $\widehat{f}_n$ from the training data.
- Need to specify the meaning of good.

Loss and Probabilistic Framework

Loss function

- $\ell(Y, f(\mathbf{X}))$: the goodness of the prediction of Y by $f(\mathbf{X})$.
- Prediction loss: $\ell(Y, f(\mathbf{X})) = \mathbf{1}_{Y \neq f(\mathbf{X})}$.
- Quadratic loss: $\ell(Y, \mathbf{X}) = \|Y - f(\mathbf{X})\|_2^2$.

Risk function

- Risk measured as the average loss:

$$\mathcal{R}(f) = \mathbb{E}[\ell(Y, f(\mathbf{X}))] \,.$$

- Prediction loss: $\mathbb{E}[\ell(Y, f(\mathbf{X}))] = \mathbb{P}(Y \neq f(\mathbf{X}))$.
- Quadratic loss: $\mathbb{E}[\ell(Y, f(\mathbf{X}))] = \mathbb{E}[\|Y - f(\mathbf{X})\|_2^2]$.

- **Beware:** As $\widehat{f_n}$ depends on $\mathcal{D}_n$, $\mathcal{R}(\widehat{f_n})$ is a random variable!

Binary classification

Bayes classifier

The Bayes classifier $g^\star$ is defined as:

$$g^\star(X) = \begin{cases} 1 & \text{if} \quad \mathbb{P}\left(Y = 1|X\right) > \mathbb{P}\left(Y = 0|X\right), \\ 0 & \text{otherwise.} \end{cases}$$

Equivalently,

$$g^\star(X) = \begin{cases} 1 & \text{if} \quad \mathbb{P}\left(Y = 1|X\right) > 1/2, \\ 0 & \text{otherwise,} \end{cases}$$

Lemma

For any classification rule $g : \mathbb{R}^d \to \{0, 1\}$, one has

$$\mathcal{R}(g^\star) \leqslant \mathcal{R}(g).$$

How to model the conditional law of Y?

In practice we do not know the conditional law of Y given X.
Several solutions to overcome this issue.

Fully parametric modeling.
Estimate the law of (X, Y) and use the **Bayes formula** to deduce an estimate of the conditional law of Y: *LDA/QDA, Naive Bayes...*

Parametric conditional modeling.
Estimate the conditional law of Y by a **parametric** law: *linear regression, logistic regression, Feed Forward Neural Networks...*

Nonparametric conditional modeling.
Estimate the conditional law of Y by a **non parametric** estimate: *kernel methods, nearest neighbors...*

Summary

Recall the linear model

- In regression with $\mathcal{X} = \mathbb{R}^d$, the linear model is the parametric reference model.
- This model makes the assumption that the regression function is linear: for $1 \leqslant i \leqslant n$

$$Y = X^\top \beta^\star + \varepsilon,$$

with

$$\mathbb{E}[\varepsilon | X] = 0 \quad \text{and} \quad \mathbb{V}[\varepsilon | X] = \sigma^2.$$

- Here, estimating the regression function is equivalent to estimate $\beta^\star \in \mathbb{R}^d$.

Finite dimensional parametric model

ERM with quadratic loss

- The least squares estimates, i.e. the optimal solution of

$$\min_{\beta \in \mathbb{R}^d} \widehat{\mathcal{R}}_n(\beta) = \min_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2$$

with $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times d}$ is given, if X has full rank, by

$$\widehat{\beta}_n = (X^\top X)^{-1} X^\top Y.$$

- $m^\star : x \mapsto x^\top \beta^*$ is estimated by $\widehat{m}_n : x \mapsto x^\top \widehat{\beta}_n$.
- Under some technical assumptions (see lectures of past year)

$$\mathbb{E}[\widehat{\beta}_n] = \beta^\star \quad \text{and} \quad \mathbb{V}(\widehat{\beta}_n) = \sigma^2 (X^\top X)^{-1}.$$

- We deduce that, see `sylvainlc.github.io`,

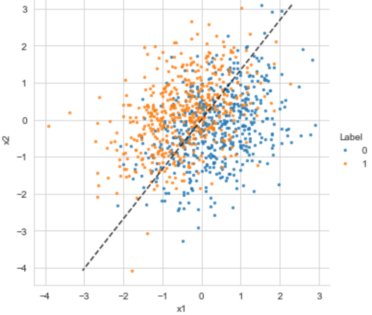
$$\mathbb{E}\left[\|\widehat{\beta}_n - \beta\|_2^2\right] = O\left(\frac{1}{n}\right) \quad \text{and} \quad \mathbb{E}\left[(\hat{m}_n(x) - m^\star(x))^2\right] = O\left(\frac{1}{n}\right)$$

Fully parametric modeling - Discriminant Analysis

In the LDA case, the classification rule is of the form:

$$g^*(x) = 1 \Leftrightarrow \langle w, x \rangle + b \geqslant 0 \,,$$

where w and b depends on the model parameters.



$\rightharpoondown$ Relax the Gaussian assumption ? (logistic model, SVM).
$\rightharpoondown$ Design nonlinear classification rules ? (kernels, neural networks).

The logistic model

- One of the most widely used classification algorithm.
- It models the distribution of Y given X. For $y \in \{0, 1\}$

$$\mathbb{P}(Y = 1|X) = \sigma\left(X^\top w + b\right)$$

where $w \in \mathbb{R}^d$ is a vector of model weights and $b \in \mathbb{R}$ is the intercept, and where $\sigma : z \mapsto (1 + \mathrm{e}^{-z})$ is the sigmoid function:

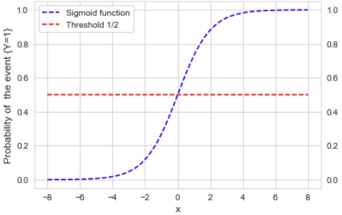


Figure: The sigmoid function

Some comments on the logistic model

- The sigmoid is a modelling choice to map $\mathbb{R} \to [0, 1]$ (to model a probability).
- We could also consider

$$\mathbb{P}(Y = 1|X) = F\left(X^\top w + b\right)$$

for any distribution function F.

- Another popular choice is the Gaussian distribution

$$F(z) = \mathbb{P}(\mathcal{N}(0, 1) \leqslant z),$$

which leads to another loss called probit.

The logistic model

- In the case of the sigmoid, one has

$$\mathbb{P}(Y = 1|X) = \frac{\exp(b + w^\top X)}{1 + \exp(b + w^\top X)} = \frac{1}{1 + \exp(-(b + w^\top X))}$$

$$\mathbb{P}(Y = 0|X) = \frac{1}{1 + \exp(b + w^\top X)}$$

- However, the sigmoid choice has the following nice interpretation: an easy computation leads to

$$\log\left(\frac{\mathbb{P}(Y = 1|X)}{\mathbb{P}(Y = 0|X)}\right) = X^\top w + b.$$

- This quantity is called the log-odd ratio.

The logistic model

- Therefore, this model makes the assumption that (the logit transformation of) the probability $p(X) = \mathbb{P}(Y = 1|X)$ is linear:

$$\mathrm{logit}(p(X)) := \log\left(\frac{p(X)}{1 - p(X)}\right) = X^\top w + b.$$

- Note that

$$\mathbb{P}(Y = 1|X) \geqslant \mathbb{P}(Y = 0|X)$$

if and only if

$$X^\top w + b \geqslant 0.$$

This is a linear classification rule, linear w.r.t. the considered features x!

The logistic "regression"

Theorem

Consider that the logit-transformation is linear with parameters $(b^\star, w^\star)$*:*

$$\mathrm{logit}(p(X)) := \log\left(\frac{p(X)}{1-p(X)}\right) = f^\star(X) = X^\top w^\star + b^\star.$$

Then $f^\star$ is a minimizer of the risk: $f \mapsto \mathbb{E}\left[\log\left(1+\exp(-Yf(X))\right)\right]$ *over all affine functions and*

$$g^\star : x \in \mathbb{R}^d \mapsto \begin{cases} 1 & \textit{if } f^\star(x) > 0 \\ 0 & \textit{otherwise} \end{cases}$$

is a Bayes classifier.

Estimation of w and b

Parametric model for the conditional law of Y given X: $\mathbb{P}_{w,b}(Y|X)$.

Compute estimators $\hat{w}$ and $\hat{b}$ by maximum likelihood estimation.

Or equivalently, minimize the minus log-likelihood:

$$(w, b) \mapsto -n^{-1} \log \mathbb{P}_{w,b}(Y_{1:n}|X_{1:n}).$$

More generally, when a model is used

$$\text{Goodness-of-fit} = \text{-log likelihood}$$

The log function is used mainly since averages are easier to study (and compute) than products.

Logistic regression - likelihood function

$\rightharpoondown$ $\{(X_i, Y_i)\}_{1\leqslant i\leqslant n}$ are i.i.d. with the same distribution as (X, Y).

Likelihood:

$$\prod_{i=1}^{n} \mathbb{P}_{w,b}(Y_i|X_i) = \prod_{i=1}^{n} \sigma(\langle w, X_i\rangle + b)^{Y_i}\big(1 - \sigma(\langle w, X_i\rangle + b)\big)^{1-Y_i},$$
$$= \prod_{i=1}^{n} \sigma(\langle w, x_i\rangle + b)^{Y_i}\sigma(-\langle w, X_i\rangle - b)^{1-Y_i}$$

and the normalized negative loglikelihood is written

$$f(w, b) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, \langle w, X_i\rangle + b)\,.$$

Logistic regression - likelihood function

Compute $\hat{w}_n$ and $\hat{b}_n$ as follows:

$$(\hat{w}_n, \hat{b}_n) \in \mathrm{argmin}_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \left(-Y_i (X_i^\top w + b) + \log(1 + e^{X_i^\top w + b}) \right) .$$

$\rightharpoondown$ It is an **average of losses**, one for each sample point.
$\rightharpoondown$ It is a convex and smooth problem.

Using the **logistic loss** function

$$\ell : (y, y') \mapsto \log(1 + e^{-yy'})$$

yields

$$(\hat{w}_n, \hat{b}_n) \in \mathrm{argmin}_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, \langle w, X_i \rangle + b) .$$

Maximum likelihood estimate

Assume for now that the intercept is 0. Then, the likelihood is,

$$L_n(w) = \prod_{i=1}^{n} \left(\frac{e^{X_i^T w}}{1 + e^{X_i^T w}} \right)^{Y_i} \left(\frac{1}{1 + e^{X_i^T w}} \right)^{1-Y_i} = \prod_{i=1}^{n} \left(\frac{e^{X_i^T w Y_i}}{1 + e^{X_i^T w}} \right) .$$

And the negative log-likelihood is

$$\ell_n(w) = -\log(L_n(w)) = \sum_{i=1}^{n} \left(-Y_i X_i^T w + \log(1 + e^{X_i^T w}) \right) .$$

Derivatives

$$\begin{aligned} \frac{\partial \left(\log(L_n(w)) \right)}{\partial w_j} &= \sum_{i=1}^{n} \left(Y_i X_{ij} - \frac{x_{ij} e^{X_i^T w}}{(1 + e^{X_i^T w})} \right) \\ &= \sum_{i=1}^{n} X_{ij} \left(Y_i - \sigma(\langle w, X_i \rangle) \right) . \end{aligned}$$

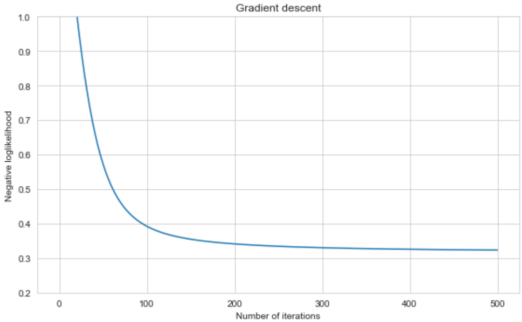
Maximum likelihood estimate

The negative loglikelihood

$$\ell_n(w) = -\log(L_n(w)) = \sum_{i=1}^{n} \left(-Y_i X_i^T w + \log(1 + e^{X_i^T w})\right) .$$

is minimized using a gradient descent algorithm.
Starting with an initial estimate $w^{(0)}$, for all $k \geqslant 1$, set

$$w^{(k)} = w^{(k-1)} - \eta_k \nabla \ell_n(w^{(k-1)}) .$$



The gradient of the negative loglikelihood is,

$$\nabla \ell_n(w) = -\sum_{i=1}^{n} Y_i X_i + \sum_{i=1}^{n} \frac{\exp(\langle X_i, w\rangle)}{1+\exp(\langle X_i, w\rangle)} X_i \,.$$

On the other hand, for all $1 \leqslant i \leqslant n$ and all $1 \leqslant j \leqslant d$,

$$\partial_j \left(\frac{\exp(\langle X_i, w\rangle)}{1+\exp(\langle X_i, w\rangle)} X_i \right) = \frac{\exp(\langle X_i, w\rangle)}{(1+\exp(\langle X_i, w\rangle))^2} X_{ij} X_i \,,$$

where X_{ij} is the jth component of X_i.

Then, the Hessian matrix is

$$H_n(w) = \sum_{i=1}^{n} \frac{\exp(\langle X_i, w\rangle)}{(1+\exp(\langle X_i, w\rangle))^2} X_i X_i^T \,.$$

Asymptotic properties

Assumptions

→ $\widehat{w}_n \to w^*$ almost surely.

→ There exists a continuous and nonsingular function H such that $n^{-1}H_n(w)$ converges to $H(w)$, uniformly in a ball around w^*.

For all $t \in \mathbb{R}^d$, using a Taylor expansion,

$$\mathbb{E}\left[\exp\left(-\frac{1}{\sqrt{n}}\langle t, \nabla \ell_n(w^\star)\rangle\right)\right] \to_{n\to\infty} \exp\left(\frac{1}{2}t^T H(w^\star)t\right).$$

Therefore,

$$-\nabla \ell_n(w^\star)/\sqrt{n} \Rightarrow \mathcal{N}(0, H(w^\star)).$$

On the other hand, by Slutsky lemma,

$$\sqrt{n}(\widehat{w}_n - w^\star) \Rightarrow \mathcal{N}(0, H(w^\star)^{-1}).$$

Confidence interval

$\rightharpoondown$ $\sqrt{n}(\widehat{w}_j - w_j^\star)$ converges in distribution to a centered Gaussian random variable with variance $(H(w^\star)^{-1})_{jj}$.

Almost surely, $\widehat{\sigma}^2_{n,j} = (nH_n(\widehat{w}_n)^{-1})_{jj} \to_{n\to\infty} (H(w^\star)^{-1})_{jj}$.
Then,

$$\sqrt{\frac{n}{\widehat{\sigma}^2_{n,j}}}(\widehat{w}_{n,j} - \beta_j^\star) \to_{n\to\infty} \mathcal{N}(0,1)\,.$$

An asymptotic confidence interval $\mathcal{I}_{n,\alpha}$ of level $1-\alpha$ is then

$$\mathcal{I}_{n,\alpha} = \left[\widehat{w}_{n,j} - z_{1-\alpha/2}\sqrt{\frac{\widehat{\sigma}^2_{n,j}}{n}}\,,\ \widehat{\beta}_{n,j} + z_{1-\alpha/2}\sqrt{\frac{\widehat{\sigma}^2_{n,j}}{n}}\right]\,,$$

where $z_{1-\alpha/2}$ is the quantile of order $1-\alpha/2$ of $\mathcal{N}(0,1)$.