Machine Learning

Linear Models

Fabio Vandin

November 6^{th}, 2023

Logistic Regression

Learn a function h from $\mathbb{R}^d$ to $[0, 1]$.

What can this be used for?

Classification!

Example: binary classification ($\mathcal{Y} = \{-1, 1\}$) - $h(\mathbf{x})$ = *probability* that label of $\mathbf{x}$ is 1.

For simplicity of presentation, we consider binary classification with $\mathcal{Y} = \{-1, 1\}$, but similar considerations apply for multiclass classification.

Logistic Regression: Model

Hypothesis class $\mathcal{H}$: $\phi_{\text{sig}} \circ L_d$, where $\phi_{\text{sig}} : \mathbb{R} \to [0,1]$ is *sigmoid function*

"linear models" $(\langle \vec{w}, \vec{x} \rangle)$

Sigmoid function = "S-shaped" function

Logistic Regression: Model

Hypothesis class $\mathcal{H}$: $\phi_{\text{sig}} \circ L_d$, where $\phi_{\text{sig}} : \mathbb{R} \to [0,1]$ is *sigmoid function*

Sigmoid function = "S-shaped" function

For logistic regression, the sigmoid ϕ_{sig} used is the *logistic regression*:

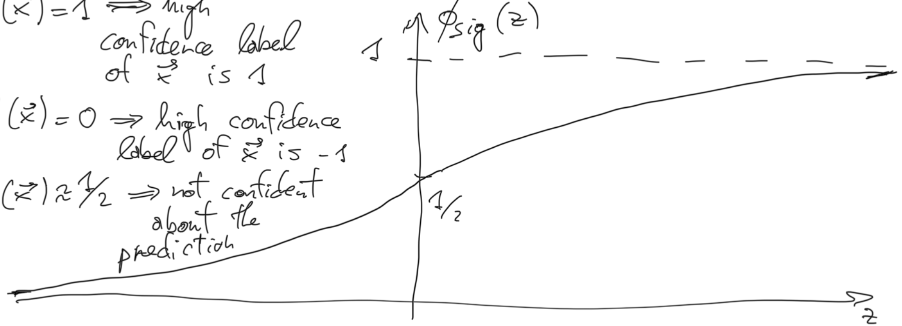
"output" of the linear model $= \langle \vec{w}, \vec{x} \rangle$

$$\phi_{\text{sig}}(z) = \frac{1}{1+e^{-z}}$$

$h(\vec{x}) = 1 \Rightarrow$ high confidence label of $\vec{x}$ is 1

$h(\vec{x}) = 0 \Rightarrow$ high confidence label of $\vec{x}$ is -1

$h(\vec{x}) \approx 1/2 \Rightarrow$ not confident about the prediction



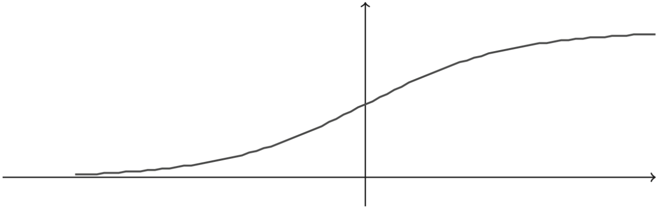
Logistic Regression: Model

Hypothesis class $\mathcal{H}$: $\phi_{\text{sig}} \circ L_d$, where $\phi_{\text{sig}} : \mathbb{R} \to [0, 1]$ is *sigmoid function*

Sigmoid function = "S-shaped" function

For logistic regression, the sigmoid ϕ_{sig} used is the *logistic regression*:

$$\phi_{\text{sig}}(z) = \frac{1}{1 + e^{-z}}$$



Therefore

parameters of the model/hypothesis

$$H_{\text{sig}} = \phi_{\text{sig}} \circ L_d = \{\mathbf{x} \to \phi_{\text{sig}}(\langle \mathbf{w}, \mathbf{x} \rangle) : \mathbf{w} \in \mathbb{R}^{d+1}\}$$

and $h_{\mathbf{w}}(\mathbf{x}) \in H_{\text{sig}}$ is:

$$h_{\mathbf{w}}(\mathbf{x}) = \frac{1}{1 + e^{-\langle \mathbf{w}, \mathbf{x} \rangle}}$$

Therefore

$$H_{\text{sig}} = \phi_{\text{sig}} \circ L_d = \{\mathbf{x} \to \phi_{\text{sig}}(\langle \mathbf{w}, \mathbf{x} \rangle) : \mathbf{w} \in \mathbb{R}^{d+1}\}$$

and $h_{\mathbf{w}}(\mathbf{x}) \in H_{\text{sig}}$ is:

$$h_{\mathbf{w}}(\mathbf{x}) = \frac{1}{1 + e^{-\langle \mathbf{w}, \mathbf{x} \rangle}}$$

Main difference with binary classification with halfspaces: when $\langle \mathbf{w}, \mathbf{x} \rangle \approx 0$

- halfspace prediction is deterministically 1 or -1
- $\phi_{\text{sig}}(\langle \mathbf{w}, \mathbf{x} \rangle) \approx 1/2 \Rightarrow$ uncertainty in predicted label

Loss Function

Need to define how bad it is to predict $h_{\mathbf{w}}(\mathbf{x}) \in [0, 1]$ given that true label is $y = \pm 1$

probability that the label is 1

What do we want?

- if $y = +1 \Rightarrow h_{\vec{w}}(\vec{x})$ large
- if $y = -1 \Rightarrow h_{\vec{w}}(\vec{x})$ small

$\Rightarrow 1 - h_{\vec{w}}(\vec{x})$ large

Loss Function

Need to define how bad it is to predict $h_{\mathbf{w}}(\mathbf{x}) \in [0, 1]$ given that true label is $y = \pm 1$

Desiderata

- $h_{\mathbf{w}}(\mathbf{x})$ "large" if $y = 1$
- $1 - h_{\mathbf{w}}(\mathbf{x})$ "large" if $y = -1$

Note that

$$
\begin{aligned}
1 - h_{\mathbf{w}}(\mathbf{x}) &= 1 - \frac{1}{1 + e^{-\langle \mathbf{w}, \mathbf{x} \rangle}} \\
&= \frac{e^{-\langle \mathbf{w}, \mathbf{x} \rangle}}{1 + e^{-\langle \mathbf{w}, \mathbf{x} \rangle}} \cdot \frac{e^{\langle \vec{w}, \vec{x} \rangle}}{e^{\langle \vec{w}, \vec{x} \rangle}} \\
&= \frac{1}{1 + e^{\langle \mathbf{w}, \mathbf{x} \rangle}}
\end{aligned}
$$

Then *reasonable* loss function: increases monotonically with

$$\frac{1}{1 + e^{y\langle \mathbf{w}, \mathbf{x} \rangle}}$$

$\Rightarrow$ *reasonable* loss function: increases monotonically with

$$1 + e^{-y\langle \mathbf{w}, \mathbf{x} \rangle}$$

Loss function for logistic regression:

$$\ell(h_{\mathbf{w}}, (\mathbf{x}, y)) = \log\left(1 + e^{-y\langle \mathbf{w}, \mathbf{x} \rangle}\right)$$

Therefore, given training set $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m))$ the ERM problem for logistic regression is:

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\arg\min} \overbrace{\frac{1}{m}\sum_{i=1}^{m} \underbrace{\log\left(1 + e^{-y_i\langle \mathbf{w},\mathbf{x}_i\rangle}\right)}_{\ell(h_{\vec{w}}, (\vec{x}_i, y_i))}}^{L_S(h_{\vec{w}})}$$

Therefore, given training set $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m))$ the ERM problem for logistic regression is:

$$\arg\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{m}\sum_{i=1}^{m} \log\left(1 + e^{-y_i\langle \mathbf{w}, \mathbf{x}_i\rangle}\right)$$

Notes: logistic loss function is a *convex function* $\Rightarrow$ ERM problem can be solved efficiently

Definition may look a bit arbitrary: actually, ERM formulation is the same as the one arising from *Maximum Likelihood Estimation*

Maximum Likelihood Estimation (MLE) [UML, 24.1]

MLE is a statistical approach for finding the parameters that maximize the joint probability of a given dataset *assuming a specific parametric probability function*.

Note: MLE essentially assumes a *generative model* for the data

General approach:

1. given training set $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m))$, assume each $(\mathbf{x}_i, y_i)$ is i.i.d. from some probability distribution of parameters θ
2. consider $\mathbb{P}[S|\theta]$ (likelihood of data given parameters)
3. *log likelihood*: $L(S; \theta) = \log(\mathbb{P}[S|\theta])$
4. *maximum likelihood estimator*: $\hat{\theta} = \arg\max_{\theta} L(S; \theta)$

Logistic Regression and MLE

Assuming $\mathbf{x}_1, \ldots, \mathbf{x}_m$ are fixed, the probability that $\mathbf{x}_i$ has label $y_i = 1$ is

$$h_{\mathbf{w}}(\mathbf{x}_i) = \frac{1}{1 + e^{-\langle \mathbf{w}, \mathbf{x}_i \rangle}}$$

parameters of the probability "function"

Logistic Regression and MLE

Assuming $\mathbf{x}_1, \ldots, \mathbf{x}_m$ are fixed, the probability that $\mathbf{x}_i$ has label $y_i = 1$ is

$$h_{\mathbf{w}}(\mathbf{x}_i) = \frac{1}{1 + e^{-\langle \mathbf{w}, \mathbf{x}_i \rangle}}$$

while the probability that $\mathbf{x}_i$ has label $y_i = -1$ is

$$(1 - h_{\mathbf{w}}(\mathbf{x}_i)) = \frac{1}{1 + e^{\langle \mathbf{w}, \mathbf{x}_i \rangle}}$$

parameters

For each i, the probability that $\vec{x}_i$ has label y_i is:

$$\frac{1}{1 + e^{-y_i \langle \vec{w}, \vec{x}_i \rangle}}$$

Logistic Regression and MLE

Assuming $\mathbf{x}_1, \ldots, \mathbf{x}_m$ are fixed, the probability that $\mathbf{x}_i$ has label $y_i = 1$ is

$$h_{\mathbf{w}}(\mathbf{x}_i) = \frac{1}{1 + e^{-\langle \mathbf{w}, \mathbf{x}_i \rangle}}$$

while the probability that $\mathbf{x}_i$ has label $y_i = -1$ is

$$(1 - h_{\mathbf{w}}(\mathbf{x}_i)) = \frac{1}{1 + e^{\langle \mathbf{w}, \mathbf{x}_i \rangle}}$$

Then the likelihood for training set S is:

$$\prod_{i=1}^{m} \underbrace{\left(\frac{1}{1 + e^{-y_i \langle \mathbf{w}, \mathbf{x}_i \rangle}} \right)}_{\Pr[\vec{x}_i \text{ has label } y_i \mid \text{parameters } \vec{w}]}$$

elements of
S are i.i.d.

Therefore the log likelihood is:

$$\log\left(\prod_{i=1}^{m} \frac{1}{1+e^{-y_i\langle\vec{w},\vec{x}_i\rangle}}\right)$$

$$= \sum_{i=1}^{m} \log\left(\frac{1}{1+e^{-y_i\langle\vec{w},\vec{x}_i\rangle}}\right)$$

$$= \sum_{i=1}^{m}\left(\underbrace{\log 1}_{=0} - \log\left(1+e^{-y_i\langle\vec{w},\vec{x}_i\rangle}\right)\right)$$

$$= -\sum_{i=1}^{m} \log\left(1+e^{-y_i\langle\vec{w},\vec{x}_i\rangle}\right)$$

Therefore the log likelihood is:

$$-\sum_{i=1}^{m} \log\left(1 + e^{-y_i \langle \mathbf{w}, \mathbf{x}_i \rangle}\right)$$

And note that the maximum likelihood estimator for $\mathbf{w}$ is:

$$\arg\max_{\mathbf{w} \in \mathbb{R}^d} -\sum_{i=1}^{m} \log\left(1 + e^{-y_i \langle \mathbf{w}, \mathbf{x}_i \rangle}\right) = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \sum_{i=1}^{m} \log\left(1 + e^{-y_i \langle \mathbf{w}, \mathbf{x}_i \rangle}\right)$$

Therefore the log likelihood is:

$$-\sum_{i=1}^{m} \log\left(1 + e^{-y_i \langle \mathbf{w}, \mathbf{x}_i \rangle}\right)$$

And note that the maximum likelihood estimator for $\mathbf{w}$ is:

$$\arg\max_{\mathbf{w}\in\mathbb{R}^d} -\sum_{i=1}^{m} \log\left(1 + e^{-y_i \langle \mathbf{w}, \mathbf{x}_i \rangle}\right) = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \sum_{i=1}^{m} \log\left(1 + e^{-y_i \langle \mathbf{w}, \mathbf{x}_i \rangle}\right)$$

$\Rightarrow$ MLE solution is equivalent to ERM solution!

Bibliography

[UML] Chapter 9:

- no 9.1.1