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 model

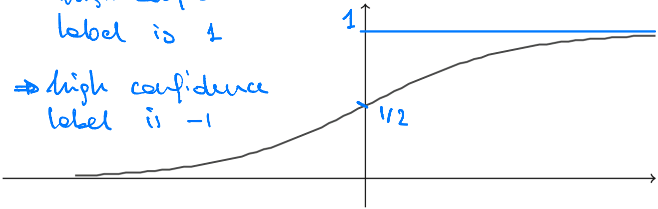
Sigmoid function = "S-shaped" function

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

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

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

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



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

Therefore

Bias

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

Linearity

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 $y = \pm 1$

Desiderata

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

Note that

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

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:

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

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

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

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:

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

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 **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!

Machine Learning

Uniform Convergence

Fabio Vandin

November 6^{th}, 2023

When is an Hypothesis Class PAC Learnable?

Previously seen result: for binary classification with

- realizability assumption
- 0-1 loss

any finite hypothesis class is PAC learnable by ERM.

What about the more general PAC learning model we have seen last? Recall the (agnostic) PAC learnability for general loss:

Definition

A hypothesis class $\mathcal{H}$ is *agnostic PAC learnable* with respect to a set Z and a loss function $\ell : \mathcal{H} \times Z \rightarrow \mathbb{R}_+$ if there exist a function $m_{\mathcal{H}}$: $(0,1)^2 \rightarrow \mathbb{N}$ and a learning algorithm such that for every $\delta, \varepsilon \in (0,1)$, for every distribution $\mathcal{D}$ over Z, when running the learning algorithm on $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$ the algorithm returns a hypothesis h such that, with probability $\geq 1 - \delta$ (over the choice of the m training examples):

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon$$

where $L_{\mathcal{D}}(h) = \mathbb{E}_{z \sim \mathcal{D}}[\ell(h, z)]$

Uniform Convergence and Learnability

Uniform convergence: the empirical risks (training error) of *all* members of $\mathcal{H}$ are good approximations of their true risk (generalization error).

Definition

A training set S is called ε*-representative* (w.r.t. domain Z, hypothesis class $\mathcal{H}$, loss function ℓ, and distribution $\mathcal{D}$) if

$$\forall h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \leq \varepsilon$$

Proposition

Assume that training set S is $\frac{\varepsilon}{2}$-representative (w.r.t. domain Z, hypothesis class $\mathcal{H}$, loss function ℓ, and distribution $\mathcal{D}$). Then, any output of $\mathrm{ERM}_{\mathcal{H}}(S)$ (i.e., any $h_S \in \arg\min_{h \in \mathcal{H}} L_S(h)$) satisfies

$$L_{\mathcal{D}}(h_S) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon$$

Proof.

For every $h \in \mathcal{H}$:

$$\begin{aligned} L_{\mathcal{D}}(h_S) &\leq L_S(h_S) + \frac{\varepsilon}{2} \\ &\leq L_S(h) + \frac{\varepsilon}{2} \\ &\leq L_{\mathcal{D}}(h) + \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \\ &= L_{\mathcal{D}}(h) + \varepsilon \end{aligned}$$

□

Uniform convergence depends on training set: when do we have uniform convergence?

Definition

A hypothesis class $\mathcal{H}$ has the *uniform convergence property* (w.r.t. a domain Z and a loss function ℓ) if there exists a function $m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$ such that for every $\varepsilon, \delta \in (0,1)$ and for every probability distribution $\mathcal{D}$ over Z, if S is a sample of $m \geq m_{\mathcal{H}}^{UC}(\varepsilon, \delta)$ i.i.d. examples drawn from $\mathcal{D}$, then with probability $\geq 1 - \delta$, S is ε-representative.

Proposition

If a class $\mathcal{H}$ has the uniform convergence property with a function $m_{\mathcal{H}}^{UC}$ then the class is agnostically PAC learnable with the sample complexity $m_{\mathcal{H}}(\varepsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\varepsilon/2, \delta)$. Furthermore, in that case the $\mathrm{ERM}_{\mathcal{H}}$ paradigm is a successful agnostic PAC learner for $\mathcal{H}$.

What classes of hypotheses have uniform convergence?

Finite Classes are Agnostic PAC Learnable

We prove that finite sets of hypotheses are agnostic PAC learnable under some restriction for the loss.

Proposition

Let $\mathcal{H}$ be a finite hypothesis class, let Z be a domain, and let $\ell : \mathcal{H} \times Z \to [0,1]$ be a loss function. Then:

- $\mathcal{H}$ enjoys the uniform convergence property with sample complexity

$$m_{\mathcal{H}}^{UC}(\varepsilon, \delta) \leq \left\lceil \frac{\log(2|\mathcal{H}|/\delta)}{2\varepsilon^2} \right\rceil$$

- $\mathcal{H}$ is agnostically PAC learnable using the ERM algorithm with sample complexity

$$m_{\mathcal{H}}(\varepsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\varepsilon/2, \delta) \leq \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)}{\varepsilon^2} \right\rceil$$

Idea of the proof:

1. prove that uniform convergence holds for a finite hypothesis class
2. use previous result on uniform convergence and PAC learnability

Useful tool: Hoeffding's Inequality

Hoeffding's Inequality

Let $\theta_1, \ldots, \theta_m$ be a sequence of i.i.d. random variables and assume that for all i, $\mathbb{E}[\theta_i] = \mu$ and $\mathbb{P}[a \leq \theta_i \leq b] = 1$. Then, for any $\varepsilon > 0$

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \varepsilon\right] \leq 2e^{-\frac{2m\varepsilon^2}{(b-a)^2}}$$

Proof (see also the book)

some steps

Fix $\varepsilon, \delta \in (0,1)$. We need a sample size m such that for any D, with probability $\geq 1-\delta$ (over the choice of $S = (z_1, \dots, z_m)$, $z_i = (\vec{x}_i, y_i)$), we have:

$$\forall h \in \mathcal{H} : \quad |L_S(h) - L_D(h)| \leq \varepsilon$$

That is: $D^m\left(\{S : \forall h \in \mathcal{H},\ |L_S(h) - L_D(h)| \leq \varepsilon\}\right) \geq 1-\delta$,

where $S = (z_1, \dots, z_m)$, $z_i = (\vec{x}_i, y_i)$, i.i.d. from D

Equivalently we need to show:

$$\underbrace{D(\{S : \exists h \in \mathcal{H},\ |L_S(h) - L_D(h)| \underline{\geq} \varepsilon\})}_{(*)} < \underline{\underline{\delta}}$$

We have: $\{S : \exists h \in \mathcal{H},\ |L_S(h) - L_D(h)| > \varepsilon\} = \bigcup_{h \in \mathcal{H}} \{S : |L_S(h) - L_D(h)| > \varepsilon\}$

Then: $(*) \leq \underbrace{\sum_{h \in \mathcal{H}} D^m(\{S : |L_S(h) - L_D(h)| > \varepsilon)}_{(**)}$