Logistic Regression

23 September 2024 14:27

Discriminative, Probabilistic Discriminative, Generative models — what do all of them do?

Input: $\bar{x}$
Assign to one of k classes C_k $[k = 1, 2 \ldots k]$

for linear models, decision surface is linear.

Input space is divided into DECISION REGIONS

Your classifier is essentially a $\boxed{\text{MAP FUNCTION}}$ → Approximating function

high dimension input space ⇒ low dimension label space

I/P : n dim^n
Dec. surf.: (n−1)
O/P : label

Decision surface is an (n−1) dimensional hyperplane
assuming n dimension I/P space

Discriminating function → $\boxed{\text{direct assignment of } \bar{x} \text{ to } C_k}$ → Learn $P(C_k/\bar{x})$ directly.
→ (KNN, SVM, CNN, LSTM, tree) based classifiers

$\boxed{\text{Discriminating Probabilistic Models.}}$ — Logistic regression!

Inference stage
- model conditional probability distribution $P(C_k|\bar{x})$
  model by estimating parameters (using training data)

Decision stage
use $P(C_k|\bar{x})$ to provide final decision → $\bar{x} \in C_k$.

Generative models → $\boxed{\text{model } P(\bar{x}|C_k)}$ instead of $P(C_k|\bar{x})$

Naive Bayes, GMM, HMM
unsupervised learning.

Using Bayes theorem : $\underset{\text{Posterior}}{P(C_k|\bar{x})} = \dfrac{\overset{\text{Likelihood}}{P(\bar{x}|C_k)} \cdot \overset{\text{Prior}}{P(C_k)}}{P(\bar{x})} \simeq \dfrac{P(\bar{x}, C_k)}{P(\bar{x})}$

essentially learning the $P(x, C_k)$ distrib^n

Probabilistic models → $\boxed{\text{probability distributions}}$, not absolute values in $\bar{x}$

- Generative: Naive Bayes, HMM, GMM
- Discriminative: Logistic reg.
- Graphical PGM (probabilistic graphical models)

→ Method 1: Bayesian inference (prior + observed data)
→ Method 2: Max. Likelihood estimate - MLE (distrib^n that best suits our data)

— x —

probability of an event happening = p

probability of an event not happening = 1−p.

① odds (event) $= \dfrac{p}{1-p} = \dfrac{\text{happening}}{\text{not happening}}$

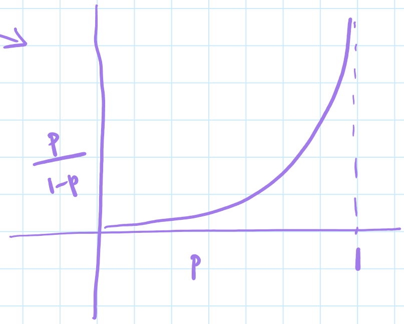
If two events: p_1, p_2
$odds(p_1) = \dfrac{p_1}{1-p_1}$ $\quad odds(p_2) = \dfrac{p_2}{1-p_2}$

① odds (event) $= \frac{p}{1-p} = \frac{\text{happening}}{\text{not happening}}$

if two events: p_1, p_2

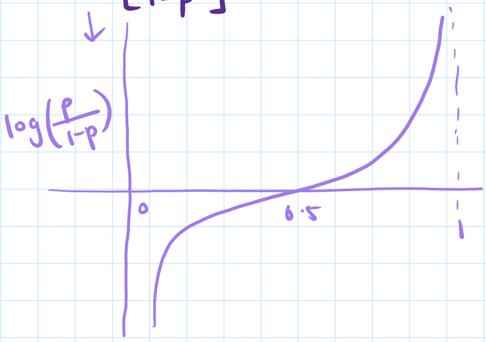
$odds(p_1) = \frac{p_1}{1-p_1}$ $\quad odds(p_2) = \frac{p_2}{1-p_2}$

odds ratio $(R) = \frac{p_1(1-p_2)}{p_2(1-p_1)}$



② Logit = log [odd]

$\ln\left[\frac{p}{1-p}\right] = \text{logit}(p)$



Estimating p:

$$\text{logit } p = w_0 + w^T x \qquad [p = p(y=1|x)]$$

$$\ln\left(\frac{p}{1-p}\right) = w_0 + w^T x$$

$$\frac{p}{1-p} = e^{w_0 + w^T x}$$

$$p = (1-p)\, e^{w_0 + w^T x}$$

$$p(1 + e^{w_0 + w^T x}) = e^{w_0 + w^T x}$$

$$p = \frac{e^{w_0 + w^T x}}{1 + e^{w_0 + w^T x}} \rightarrow \boxed{p = \frac{1}{1 + e^{-w_0 + w^T x}}}$$

from there.

Logistic regression:

" logit changes linearly w/ z when $z = w_0 + w^T x$ "

$$p(y=1|z) = \sigma(z) = \frac{1}{1 + e^{-(w_0 + w^T x)}}$$

$$\left[\frac{\partial}{\partial z}(\sigma(z)) = \sigma(z)\cdot(1 - \sigma(z))\right]$$

get target class from input z (not $\bar{x}$) where $z = w_0 + w^T x$.

Estimating parameters in logistic regression

$$p(y=1|x) = \frac{1}{1 + e^{-(w_0 + w^T x)}} \longrightarrow$$ estimate $w_0, w_1 \cdots w_n$ for n dimensional sample

minimize diff between predicted label & Actual class label.

Binary cross entropy as a loss f^n for log reg.

Bernoulli — two outcomes $< \begin{matrix} p \\ 1-p \end{matrix}$

finite, independent trials.

Namita Achyuthan

finite, independent trials.

→ Log reg follows bernoulli w/ the difference that we do not know p & have to estimate.

Say, $\{\phi_n, C_n\}$ — n samples, n classes — $c_n \in \{0,1\}$] Mathematically, to estimate $\bar{\omega}$, we look @ n samples of class 0 or 1.

for samples of label 1, $\bar{\omega}$ estimate should be so that $\prod p(x)$ for all such samples ≈ 1.

$$p(c|w) = \prod_{i=1}^{N} y_n^{c_n} (1-y_n)^{1-c_n}$$

$$\log(p(c|w)) = \sum c_n \log y_n + (1-c_n)\log(1-y_n) \rightarrow y_n = \sigma(w^T \phi_n) \text{ predicted label.}$$

Take −ve log likelihood (loss minimization)

$$\varepsilon(w) = -\log p(c|z)$$
$$= -\sum c_n \log y_n + (1-c_n)\log(1-y_n)$$

solve the maximiz^n problem now (cuz −ve log likelihood)

$$-\frac{\partial \nabla \varepsilon}{\partial w} = \frac{\partial}{\partial w} \sum c_n \log[\sigma(w^T\phi)] + (1-c_n)\log[1-\sigma(w^T\phi)]$$

Chain rule + $\frac{\partial \sigma}{\partial x} = \sigma(x) \cdot 1 - \sigma(x)$

$$-\frac{\partial \varepsilon}{\partial w} = \sum (\sigma(w^T\phi) - c)\, \phi_n$$
$$= \sum (y_n - c_n)\, \phi_n$$ — how the heck do you solve this.

$\nabla \varepsilon_w = 0$? No. $\sigma(x)$ is non linear.

Transcendental equation (can't be expressed as finite sequence of algebraic ops)

Estimate w by:
- gradient descent over likelihood fn
- Newton-Rhapson's method.

Assumptions of logreg:

a. Independent variables
b. Binary dependent variables.
c. Linear rel^n ship b/w independent var & log odds.
d. No outliers.
e. Large sample size.

Pros and Cons of Logistic Regression

The advantages of Logistic Regression are -

- Logistic regression performs better when the **data is linearly separable.**
- It **does not require too many computational resources** as it's highly interpretable.
- There is no problem scaling the input features - it **does not require tuning.**
- It is **easy to implement and train** a model using logistic regression.
- It gives a measure of how **relevant** a predictor (coefficient size) is, and its **direction of association** (positive or negative)

The disadvantages of Logistic Regression are -

- Logistic regression **fails to predict a continuous outcome.**
- Logistic regression **assumes linearity** between the predicted (dependent) variable

Pros and Cons of Logistic Regression

The advantages of Logistic Regression are -

- Logistic regression performs better when the **data is linearly separable.**
- It **does not require too many computational resources** as it's highly interpretable.
- There is no problem scaling the input features - it **does not require tuning.**
- It is **easy to implement and train** a model using logistic regression.
- It gives a measure of how **relevant** a predictor (coefficient size) is, and its **direction of association** (positive or negative)

The disadvantages of Logistic Regression are -

- Logistic regression **fails to predict a continuous outcome.**
- Logistic regression **assumes linearity** between the predicted (dependent) variable and the predictor (independent) variables.
- Logistic regression **may not be accurate if the sample size is too small.**

MLE of log reg: $F(g(x)) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$

for samples labelled '1', find β such that $\prod p(x) = 1$

for samples labelled '0', find β such that $\prod (1-p(x)) = 1$

$$L(\beta) = \prod_{s \text{ in } y_i=1} p(x_i) \times \prod_{s \text{ in } y_i=0} (1-p(x_i))$$

likelihood f^n

$$= \prod \left(p(x_i)^{y_i} \times (1-p(x_i))^{1-y_i} \right)$$

$$= \sum y_i \log(p(x_i)) + (1-y_i)\log(1-p(x_i))$$

$$= \sum y_i \log\left(\frac{1}{1+e^{-\beta x_i}}\right) + (1-y_i)\log\left(\frac{e^{-\beta x_i}}{1+e^{-\beta x_i}}\right)$$

$$= \sum y_i \left[\log\left(\frac{1}{1+e^{-\beta x_i}}\right) - \log\left(\frac{e^{-\beta x_i}}{1+e^{-\beta x_i}}\right)\right] + \log\left(\frac{e^{-\beta x_i}}{1+e^{-\beta x_i}}\right)$$

$$= \sum y_i \log(e^{\beta x_i}) + \log\left(\frac{e^{-\beta x_i}}{1+e^{-\beta x_i}} \times \frac{e^{\beta x_i}}{e^{\beta x_i}}\right)$$

$$= \sum y_i \beta x_i + \log\left(\frac{1}{1+e^{\beta x_i}}\right)$$

$$= \sum y_i \beta x_i - \log(1+e^{\beta x_i})$$