logistic regression

used for classification

training data : $\{(\vec{x}_1, y_1), \dots, (\vec{x}_n, y_n)\}$

- feature vector $\vec{x}_i \in \mathbb{R}^d$ ← how many features
- label $y_i \in \{0, 1\}$

in logistic regression, we learn from the data a probablistic model for

$Pr(Y=y \mid \vec{x})$ for $Y=0$ and $Y=1$.

(label, feature)

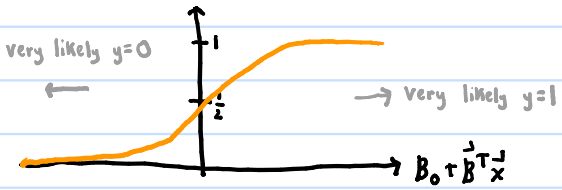
given a data pt. w./ feature vector $\vec{x}$, what is the probability that its label is $y=1$? $Pr(Y=1 \mid \vec{x})$

the logistic model: a linear model for the log odds

$$\log\left(\frac{Pr(Y=1|\vec{x})}{1-Pr(Y=1|\vec{x})}\right) = \beta_0 + \vec{\beta}^T\vec{x}$$

odds that $Y=1$ given $\vec{x}$; β_0: intercept $\in \mathbb{R}$; $\vec{\beta}$: coefficients $\vec{\beta} \in \mathbb{R}^d$

$$Pr(Y=1|\vec{x}) = \frac{e^{\beta_0 + \vec{\beta}^T\vec{x}}}{1 + e^{\beta_0 + \vec{\beta}^T\vec{x}}} \in (0,1)$$ ☺



and $Pr(Y=0|\vec{x}) = 1 - Pr(Y=1|\vec{x})$.

how to learn $\beta_0, \vec{\beta}$ from data?

under this probablistic model, write the likelihood, the probability of seeing the data given the probabalistic model & its parameters, then choose the parameters that maximize the likelihood.

$$\ell(\beta_0, \vec{\beta}) = \prod_{i=1}^{n} Pr(Y=y_i|\vec{x}) = \prod_{i=1}^{n} Pr(Y=1|\vec{x})^{y_i}\left[1-Pr(Y=1|\vec{x})\right]^{1-y_i}$$

likelihood; assumes independent, identically distributed data

$$\log \ell = \sum_{i=1}^{n} y_i \log\left[Pr(Y=1|\vec{x})\right] + (1-y_i)\log\left[1-Pr(Y=1|\vec{x})\right]$$

$$= \sum_{i=1}^{n} y_i \log\left[\frac{e^{\beta_0+\vec{\beta}^T\vec{x}_i}}{1+e^{\beta_0+\vec{\beta}^T\vec{x}_i}}\right] + (1-y_i)\log\left[\frac{1}{1+e^{\beta_0+\vec{\beta}^T\vec{x}}}\right]$$

$$= \sum_{i=1}^{n} y_i\left(\beta_0 + \vec{\beta}^T\vec{x}_i\right) - \log\left(1+e^{\beta_0+\vec{\beta}^T\vec{x}_i}\right)$$

maximizing $\log \ell$ gives same $\beta_0, \vec{\beta}$ as maximizing ℓ. $\log \ell$ is monotonic...

$$\rightarrow \quad \nabla_{\vec{\beta}} \log \ell = \vec{0}$$ to find $\vec{\beta}$ that fits the data ("learn", "train")

$$\sum_{i=1}^{n} \vec{x}_i\left(y_i - Pr(Y=1|\vec{x})\right) = \vec{0}$$

... non-linear in $\vec{\beta}$

one option to solve for $\vec{\beta}$: the Newton-Raphson algorithm.