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

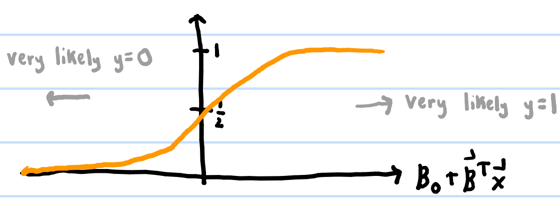
$$\Pr(\underset{\text{label}}{Y} = y \mid \underset{\text{feature}}{\vec{x}}) \quad \text{for } Y=0 \text{ and } Y=1.$$

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

$$\log\Bigg(\underbrace{\frac{\Pr(Y=1 \mid \vec{x})}{1-\Pr(Y=1 \mid \vec{x})}}_{\text{odds that } Y=1 \text{ given } \vec{x}}\Bigg) = \underset{\substack{\text{intercept} \\ \in \mathbb{R}}}{\beta_0} + \underset{\substack{\text{coefficients} \\ \vec{\beta} \in \mathbb{R}^d}}{\vec{\beta}^T} \vec{x}$$

$$\boxed{\Pr(Y=1 \mid \vec{x}) = \frac{e^{\beta_0 + \vec{\beta}^T \vec{x}}}{1 + e^{\beta_0 + \vec{\beta}^T \vec{x}}}} \in (0,1) \quad \ddot\smile$$



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

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.

$$\underset{\text{likelihood}}{\ell(\beta_0, \vec{\beta})} = \underset{\substack{\text{assumes independent,} \\ \text{identically distributed data}}}{\prod_{i=1}^{n}} \Pr(Y=y_i \mid \vec{x}) = \prod_{i=1}^{n} \Pr(Y=1 \mid \vec{x})^{y_i} [1 - \Pr(Y=1 \mid \vec{x})]^{1-y_i}$$

$$\begin{aligned}
\log \ell &= \sum_{i=1}^{n} y_i \log[\Pr(Y=1 \mid \vec{x})] + (1-y_i) \log[1 - \Pr(Y=1 \mid \vec{x})] \\
&= \sum_{i=1}^{n} y_i \log\left[\frac{e^{\beta_0 + \vec{\beta}^T \vec{x}_i}}{1 + e^{\beta_0 + \vec{\beta}^T \vec{x}_i}}\right] + (1-y_i) \log\left[\frac{1}{1 + e^{\beta_0 + \vec{\beta}^T \vec{x}}}\right] \\
&= \sum_{i=1}^{n} y_i (\beta_0 + \vec{\beta}^T \vec{x}_i) - \log(1 + e^{\beta_0 + \vec{\beta}^T \vec{x}_i})
\end{aligned}$$

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

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

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

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

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

confusion matrix

once a cutoff $P(Y=1|\vec{x})$ is decided upon for a decision boundary, the confusion matrix summarizes classification performance.

		TRUTH $y=1$	TRUTH $y=0$
PREDICTION	$y=1$	TP	FP
	$y=0$	FN	TN

TP : true positive
  # data whose true and predicted label is $y=1$

TN : true negative
  # data whose true and predicted label is $y=0$

FP : false positive
  # data whose predicted label is $y=1$, but true label is $y=0$

FN : false negative
  # data whose predicted label is $y=0$, but true label is $y=1$

receiver operator characteristic (ROC) curve

LR gives us $\Pr(Y=1|\vec{x}) \in (0,1)$.

to classify (map $\vec{x}$ to 0 or 1), we must choose a threshold p^* for the classification rule:

$$y(\vec{x}) = \begin{cases} 0 & \text{if } \Pr(Y=1|\vec{x}) \leq p^* \\ 1 & \quad\quad\quad\quad\quad\quad > p^* \end{cases}$$

- choice of p^* should reflect the costs of FP/FN's
- $p^* \to 0 \Rightarrow$ fewer FN, more FP $\}$ in the limit, all data is labeled by model as $y=1$
- $p^* \to 1 \Rightarrow$ fewer FP, more FN $\quad y=0$

the ROC curve scans all possible thresholds, p^*, and plots TPR vs. FPR

TPR: true positive rate ≡ recall ≡ sensitivity

$$= \frac{TP}{P} = \frac{TP}{TP+FN}$$

"fraction of true +'s predicted to be +'s"

total # that are truly positive

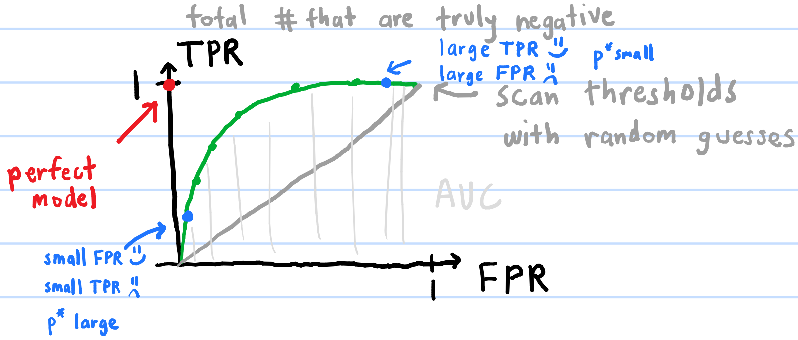
FPR: false positive rate ≡ fall out ≡ 1 − selectivity

$$= \frac{FP}{N} = \frac{FP}{TN+FP}$$

"fraction of true −'s predicted to be +'s"

total # that are truly negative

commonly used to evaluate LR model w/o imposing a p^*)



area under ROC curve (AUC):

1. randomly select a positive instance $\vec{x}_+$
2. randomly select a negative instance $\vec{x}_-$
3. $AUC = \Pr\left[\Pr(Y=1|\vec{x}_+) > \Pr(Y=1|\vec{x}_-)\right]$

i.e. AUC is probability LR properly ranks the two instances.