Logistic Regression

Classification

Aarti Singh & Geoff Gordon

Machine Learning 10-701
Feb 22, 2021

Logistic Regression

Not really regression

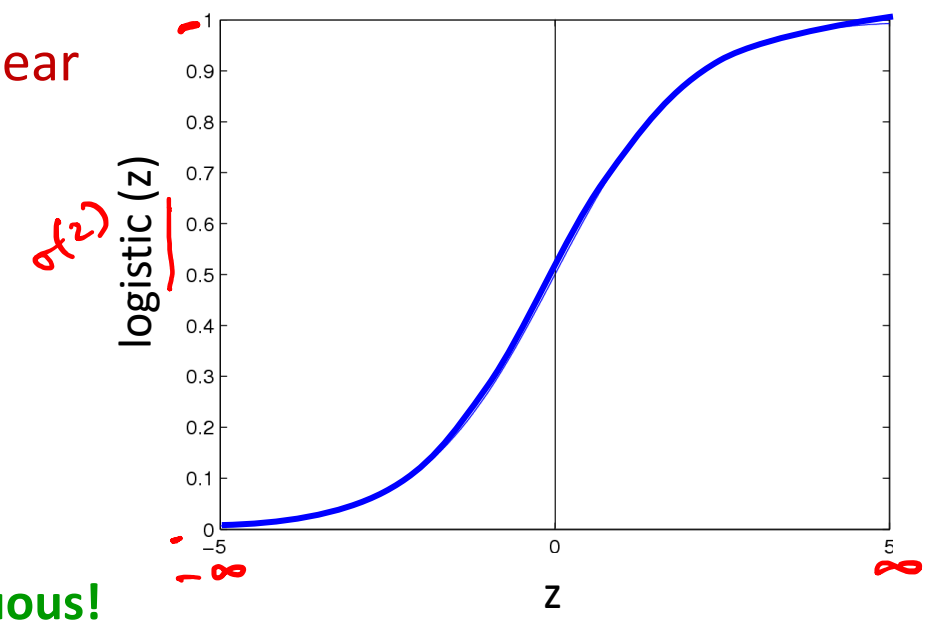
Assumes the following functional form for P(Y|X):

$$P(Y=1|X) = \frac{1}{1+\exp(-w_0 - \sum_i w_i X_i)} = \sigma\left(w_0 + \sum_i w_i X_i\right)$$

$X = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}$

Logistic function applied to a linear function of the data

Logistic function (or Sigmoid), σ(z) = : $\frac{1}{1+exp(-z)}$



Features can be discrete or continuous!

Logistic Regression is a Linear Classifier!

Assumes the following functional form for P(Y|X):

$$P(Y=1|X) = \frac{1}{1+\exp(-w_0 - \sum_i w_i X_i)}$$

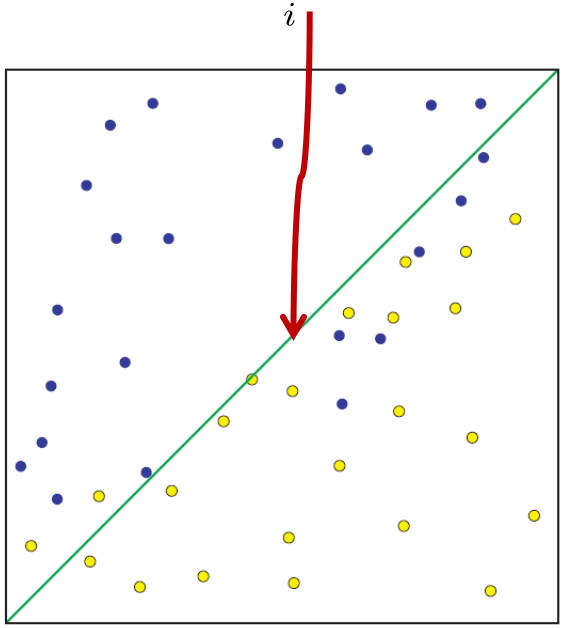
$$\Rightarrow P(Y=0|X) = \frac{e^{-(\ )}}{1+e^{-(\ )}}$$

Decision boundary:

$$\Rightarrow \frac{P(Y=1|X)}{P(Y=0|X)} = \exp(w_0 + \sum_i w_i X_i) \underset{0}{\overset{1}{\gtrless}} 1$$

$$\Rightarrow \boxed{w_0 + \sum_i w_i X_i \underset{0}{\overset{1}{\gtrless}} 0}$$

$$w_0 + \sum_i w_i X_i = 0$$



(Linear Decision Boundary)

Training Logistic Regression

How to learn the parameters $w_0, w_1, \ldots w_d$? (d features)

Training Data $\{(X^{(j)}, Y^{(j)})\}_{j=1}^{n}$ $\quad X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)})$

Maximum Likelihood Estimates

$w = (w_0 \ldots w_d)$

$$\widehat{\mathbf{w}}_{MLE} = \arg\max_{\mathbf{w}} \prod_{j=1}^{n} P(\underbrace{X^{(j)}, Y^{(j)}}_{D} \mid \mathbf{w})$$

But there is a problem …

Don't have a model for P(X) or P(X|Y) – only for P(Y|X)

Training Logistic Regression

How to learn the parameters w_0, w_1, … w_d? (d features)

Training Data $\{(X^{(j)}, Y^{(j)})\}_{j=1}^{n}$ $\quad X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)})$

Maximum (Conditional) Likelihood Estimates

$$\widehat{\mathbf{w}}_{MCLE} = \arg\max_{\mathbf{w}} \prod_{j=1}^{n} P(Y^{(j)} \mid X^{(j)}, \mathbf{w})$$

Discriminative philosophy – Don't waste effort learning P(X), focus on P(Y|X) – that's all that matters for classification!

$$\arg\max_y P(Y|X) = \arg\max_y \underbrace{P(X|Y)}_{\sim N(\ )} P(Y)$$

Expressing Conditional log Likelihood

$$P(Y=1|X,w) = \frac{1}{1+\exp(-w_0 - \sum_i w_i X_i)}$$

$$\frac{e^{-(\ )}}{1+e^{-(\ )}}$$

$$P(Y=0|X,w) = \frac{1}{1+\exp(w_0 + \sum_i w_i X_i)}$$

log

$$l(\mathbf{w}) \equiv \ln \prod_j P(y^j|\mathbf{x}^j, \mathbf{w})$$

Expressing Conditional log Likelihood

$$P(Y=1|X,w) = \frac{1}{1+\exp(-w_0 - \sum_i w_i X_i)}$$

$$P(Y=0|X,w) = \frac{1}{1+\exp(w_0 + \sum_i w_i X_i)}$$

$$\begin{aligned} l(\mathbf{w}) &\equiv \ln \prod_j P(y^j|\mathbf{x}^j, \mathbf{w}) \\ &= \sum_j \left[y^j (w_0 + \sum_i^d w_i x_i^j) - \ln(1 + exp(w_0 + \sum_i^d w_i x_i^j)) \right] \end{aligned}$$

Good news: *l*(**w**) is concave function of **w !**

Bad news: no closed-form solution to maximize *l*(**w**) HW2!

Good news: can use iterative optimization methods (gradient ascent)

That's M(C)LE. How about M(C)AP?

$$p(\mathbf{w} \mid Y, \mathbf{X}) \quad \propto \quad P(Y \mid \mathbf{X}, \mathbf{w}) p(\mathbf{w})$$

- Define priors on **w**
  - Common assumption: Normal distribution, zero mean, identity covariance
  - "Pushes" parameters towards zero

$$p(\mathbf{w}) = \prod_i \frac{1}{\kappa\sqrt{2\pi}} \quad e^{\frac{-w_i^2}{2\kappa^2}}$$

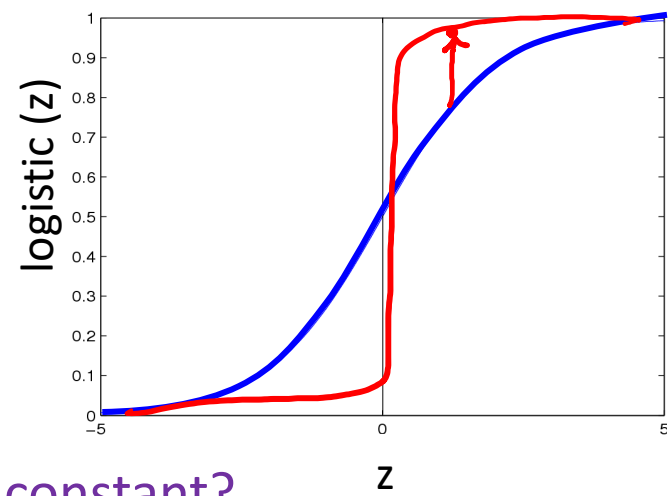
Zero-mean Gaussian prior

$w \sim N(0, \kappa^2 I)$

Logistic function (or Sigmoid):

$$\frac{1}{1 + exp(-z)}$$

$z = \sum_i w_i x_i + w_0$

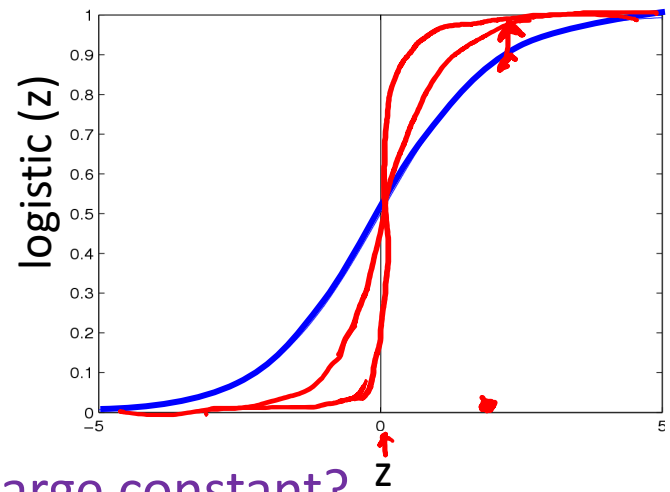


➢ What happens if we scale z by a large constant?

$z = w_0 + \sum_i w_i x_i$

Logistic function (or Sigmoid):

$$\frac{1}{1 + exp(-z)}$$



➢ Poll: What happens if we scale z by a large constant?

A) The logistic classifier decision boundary shifts towards class 1
B) The logistic classifier decision boundary remains same
C) The logistic classifier tries to separate the data perfectly
D) The logistic classifier allows more mixing of labels on each side of decision boundary

That's M(C)LE. How about M(C)AP?

$$p(\mathbf{w} \mid Y, \mathbf{X}) \quad \propto \quad P(Y \mid \mathbf{X}, \mathbf{w}) p(\mathbf{w})$$

$$p(\mathbf{w}) = \prod_i \frac{1}{\kappa\sqrt{2\pi}} \quad e^{\frac{-w_i^2}{2\kappa^2}}$$

- M(C)AP estimate

Zero-mean Gaussian prior

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \ln \left[p(\mathbf{w}) \prod_{j=1}^{n} P(y^j \mid \mathbf{x}^j, \mathbf{w}) \right]$$

$$\ln p(w) + \sum_{i=1}^{n} \ln P(y^j \mid x^j, w)$$

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \sum_{j=1}^{n} \ln P(y^j \mid \mathbf{x}^j, \mathbf{w}) - \underbrace{\sum_{i=1}^{d} \frac{w_i^2}{2\kappa^2}}_{\text{Penalizes large weights}} \quad \leftarrow \alpha \cdot \|w\|^2$$

Still concave objective!

Logistic Regression for more than 2 classes

- Logistic regression in more general case, where $Y \in \{y_1,\ldots,y_K\}$

for $k<K$

$$P(Y=y_k|X) = \frac{\exp(w_{k0} + \sum_{i=1}^{d} w_{ki}X_i)}{1 + \sum_{j=1}^{K-1} \exp(w_{j0} + \sum_{i=1}^{d} w_{ji}X_i)}$$

for $k=K$ (normalization, so no weights for this class)

$$P(Y=y_K|X) = \frac{1}{1 + \sum_{j=1}^{K-1} \exp(w_{j0} + \sum_{i=1}^{d} w_{ji}X_i)}$$

Predict $f^*(x) = \arg\max_{Y=y} P(Y=y|X=x)$

Is the decision boundary still linear?

Gaussian Naïve Bayes vs. Logistic Regression

Set of Gaussian Naïve Bayes parameters (feature variance independent of class label)

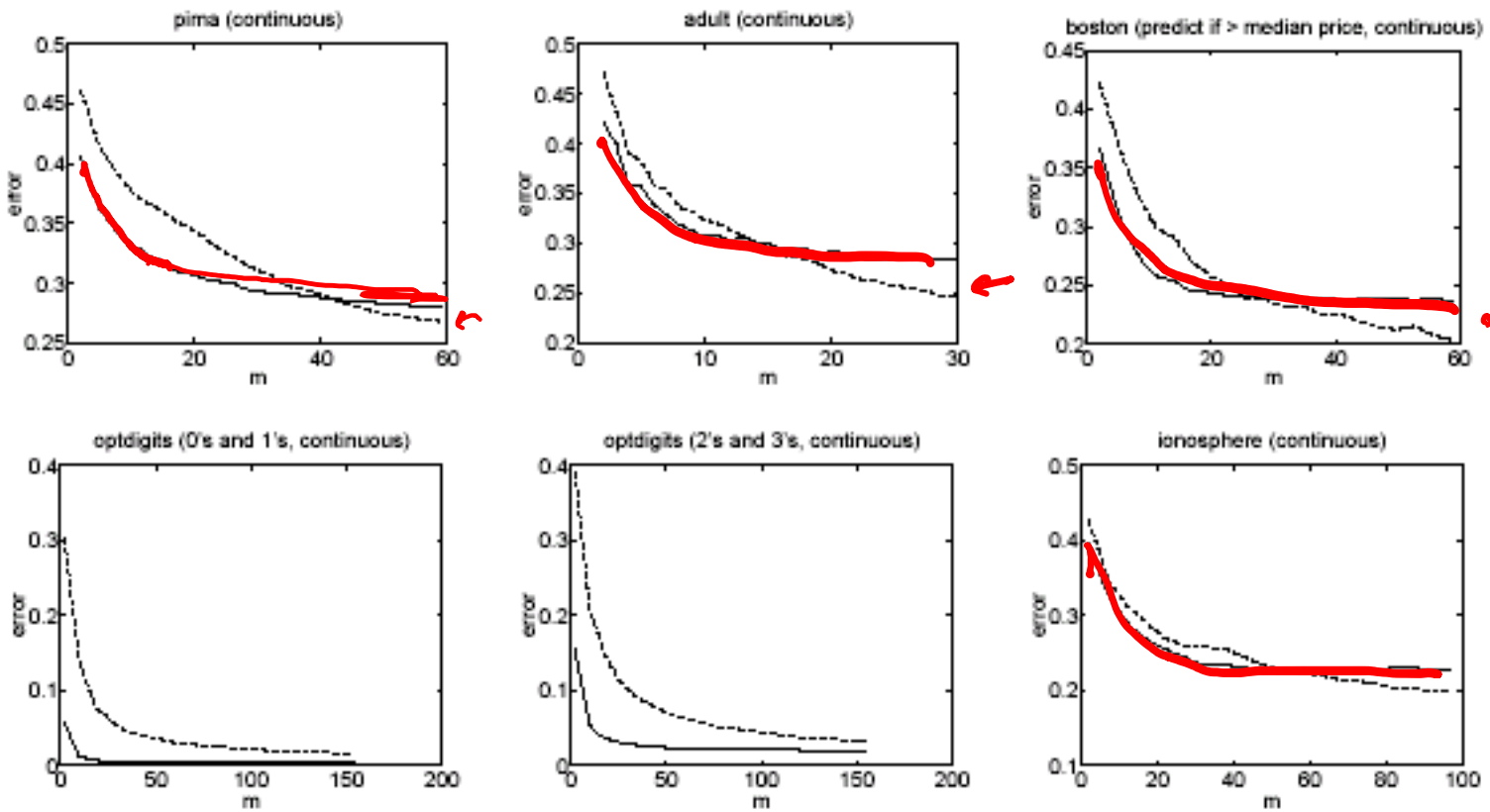
⟺

Set of Logistic Regression parameters

- Representation equivalence (both can yield linear decision boundaries)
  - **But only in a special case!!!** (GNB with class-independent variances)
  - **LR makes no assumptions about** P(**X**|Y) **in learning**!!!
  - **Optimize different functions (MLE/MCLE) or (MAP/MCAP)! Obtain different solutions**

Experimental Comparison (Ng-Jordan'01)

UCI Machine Learning Repository 15 datasets, 8 continuous features, 7 discrete features



More in Paper…

Naïve Bayes ----- Logistic Regression