CS480/680: Introduction to Machine Learning

Lecture 4: Logistic Regression

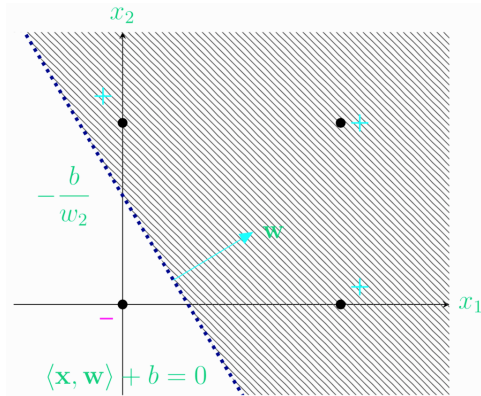
Hongyang Zhang

UNIVERSITY OF WATERLOO

Jan 23, 2024

Predicting with Confidence

- Caveat: Though called "logistic regression", it is for linear classification
- Recall that $\hat{y} = \text{sign}(\langle \mathbf{x}, \mathbf{w} \rangle)$ (assume using padding trick in this lecture)
- How confident we are about the prediction $\hat{y}$?
- Use $|\langle \mathbf{x}, \mathbf{w} \rangle|$ (margin) as an indication
  - $\langle \mathbf{x}, \mathbf{w} \rangle$ is a.k.a. logit
  - in fact was used in preceptron
  - un-normalized w.r.t. the radius of data: hard to interpret
  - need ways to normalize it into $[0, 1]$



- Better idea: learn confidence directly

Max Likelihood Estimation (MLE)

- Let $\mathcal{Y} = \{0, 1\}$; Let's directly learn confidence $p(\mathbf{x}; \mathbf{w}) := \Pr(\mathsf{Y} = 1 | \mathsf{X} = \mathbf{x})$
- Given $(\mathbf{x}_i, \mathsf{y}_i), i = 1, \ldots, n$, assume independence:

$$\Pr(\mathsf{Y}_1 = \mathsf{y}_1, \ldots, \mathsf{Y}_n = \mathsf{y}_n | \mathsf{X}_1 = \mathbf{x}_1, \ldots, \mathsf{X}_n = \mathbf{x}_n) = \prod_{i=1}^{n} \Pr(\mathsf{Y}_i = \mathsf{y}_i | \mathsf{X}_i = \mathbf{x}_i)$$

$$\overset{\mathcal{Y}=\{0,1\}}{=} \prod_{i=1}^{n} [p(\mathbf{x}_i; \mathbf{w})]^{\mathsf{y}_i} [1 - p(\mathbf{x}_i; \mathbf{w})]^{1-\mathsf{y}_i}$$

- Maximizing the likelihood:

$$\max_{\mathbf{w}} \prod_{i=1}^{n} [p(\mathbf{x}_i; \mathbf{w})]^{\mathsf{y}_i} [1 - p(\mathbf{x}_i; \mathbf{w})]^{1-\mathsf{y}_i}$$

or minimizing the minus log-likelihood:

$$\min_{\mathbf{w}} \sum_{i=1}^{n} [-\mathsf{y}_i \log p(\mathbf{x}_i; \mathbf{w}) - (1 - \mathsf{y}_i) \log(1 - p(\mathbf{x}_i; \mathbf{w}))]$$

The Logit Transform

- $p(\mathbf{x}; \mathbf{w}) : \mathcal{X} \to [0, 1]$. How to parameterize $p(\mathbf{x}; \mathbf{w})$ using $\mathbf{w}$?
- Odds Ratio $= \frac{\text{probability of event}}{\text{probability of non-event}}$

Assumption: the log of odds ratio is a linear function

Let's assume: $\log \frac{p(\mathbf{x};\mathbf{w})}{1-p(\mathbf{x};\mathbf{w})} = \langle \mathbf{x}, \mathbf{w} \rangle$.

- $\mathbf{x}$ has been padded with 1; $\mathbf{w}$ has been padded with b
- Equivalently, the Sigmoid transformation: $p(\mathbf{x}; \mathbf{w}) = \frac{1}{1+\exp(-\langle \mathbf{x}, \mathbf{w} \rangle)}$

Logistic Regression

- The minus log-likelihood that we talked before:

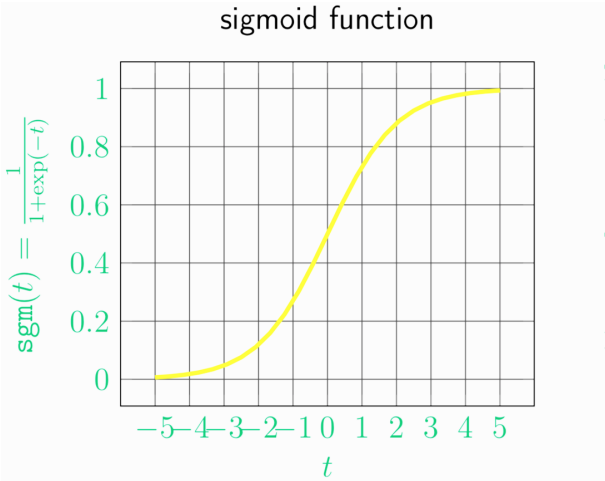
$$\min_{\mathbf{w}} \sum_{i=1}^{n} -\mathsf{y}_i \log[p(\mathbf{x}_i;\mathbf{w})] - (1-\mathsf{y}_i)\log[1-p(\mathbf{x}_i;\mathbf{w})]$$

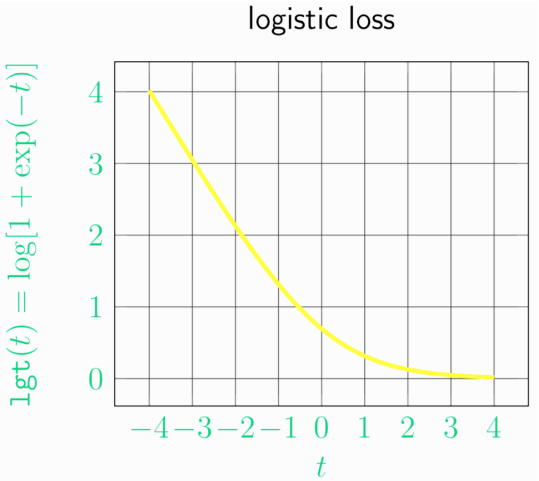
- Plug in the sigmoid parameterization $p(\mathbf{x};\mathbf{w}) = \frac{1}{1+\exp(-\langle \mathbf{x},\mathbf{w}\rangle)}$

$$\min_{\mathbf{w}} \sum_{i=1}^{n} \boxed{\log[1+\exp(-\langle \mathbf{x}_i,\mathbf{w}\rangle)] + (1-\mathsf{y}_i)\langle \mathbf{x}_i,\mathbf{w}\rangle}$$

- Note the label encoding $\mathsf{y}_i \in \{0,1\}$ (let's call this y_i^{old}); if instead, $\mathsf{y}_i^{new} \in \{\pm 1\}$, then

$$\min_{\mathbf{w}} \sum_{i=1}^{n} \underbrace{\boxed{\log[1+\exp(-\mathsf{y}_i^{new}\langle \mathbf{x}_i,\mathbf{w}\rangle)]}}_{\text{logistic loss}} \quad \text{(by transformation } \mathsf{y}_i^{old} = \frac{\mathsf{y}_i^{new}+1}{2}\text{)}$$





D. R. Cox (1958). "The Regression Analysis of Binary Sequences". Journal of the Royal Statistical Society. Series B (Methodological), vol. 20, no. 2, pp. 215–242.

Training: Solving Logistic Regression

$$\min_{\mathbf{w}} \sum_{i=1}^{n} \boxed{\log[1 + \exp(-\mathsf{y}_i \langle \mathbf{x}_i, \mathbf{w} \rangle)]}$$

- Gradient descent algorithm:

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \cdot \nabla_{\mathbf{w}} \mathsf{Loss}(\mathbf{w})$$

- We will defer the introduction of gradient descent to Lecture 8.

Prediction

$$p(\mathbf{x}; \mathbf{w}) = \texttt{sigmoid}(\langle \mathbf{x}, \mathbf{w} \rangle) = \frac{1}{1 + \exp(-\langle \mathbf{x}, \mathbf{w} \rangle)}$$

- $\hat{\mathsf{y}} = 1$ iff $p(\mathbf{x}; \mathbf{w}) = \Pr(\mathsf{Y} = 1 | \mathsf{X} = \mathbf{x}) > \frac{1}{2}$ iff $\langle \mathbf{x}, \mathbf{w} \rangle > 0$
  - $\Pr(\mathsf{Y} = 1 | \mathsf{X} = \mathbf{x}) > \frac{1}{2} \Leftrightarrow \Pr(\mathsf{Y} = 1 | \mathsf{X} = \mathbf{x}) > \Pr(\mathsf{Y} = -1 | \mathsf{X} = \mathbf{x})$
    - Therefore, $\hat{\mathsf{y}} = \operatorname{argmax}_k \Pr(\mathsf{Y} = k | \mathsf{X} = \mathbf{x})$
- Decision boundary remains to be $H := \{\mathbf{x} : \langle \mathbf{x}, \mathbf{w} \rangle = 0\}$
- Can predict $\hat{\mathsf{y}} = \operatorname{sign}(\langle \mathbf{x}, \mathbf{w} \rangle)$ as before, but now with confidence $p(\mathbf{x}; \mathbf{w})$
- An optional way to predict: check whether $p(\mathbf{x}; \mathbf{w}) > \frac{1}{2}$

More than a Classification Algorithm

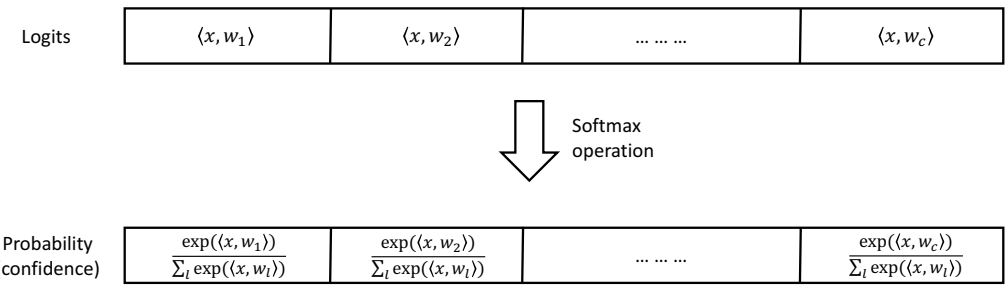
- Logistic regression estimates the posterior probability $\eta(\mathbf{x}) := \Pr(\mathsf{Y} = 1 | \mathsf{X} = \mathbf{x})$ under the linear log-odds ratio assumption
  - Confidence is meaningless if the assumption does not hold true
- Classification itself only requires comparing $\eta(\mathbf{x})$ with $\frac{1}{2}$

Multi-Class Extension

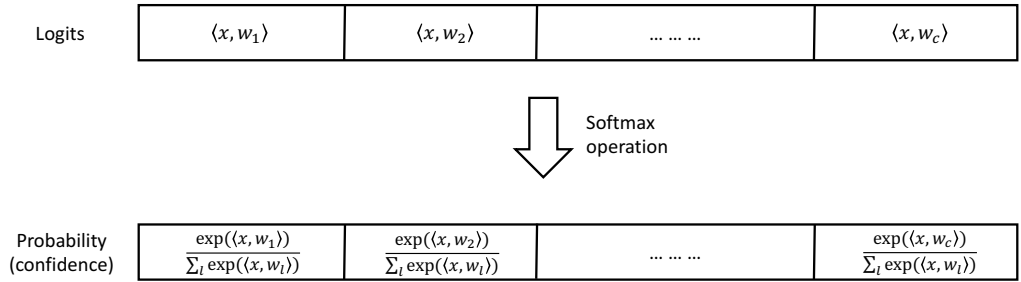
- Class y $\in \{1, \ldots, c\}$; learn $\{\mathbf{w}_1, \ldots, \mathbf{w}_c\}$ for each class
- Sigmoid $\rightarrow$ Softmax function:

$$\Pr(\mathsf{Y} = k | \mathsf{X} = \mathbf{x}; \mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_c]) = \frac{\exp(\langle \mathbf{x}, \mathbf{w}_k \rangle)}{\sum_{l=1}^{c} \exp(\langle \mathbf{x}, \mathbf{w}_l \rangle)}$$

  - map a real-valued vector to a probability vector
  - non-negative and sum to 1



Multi-Class Extension



- Training: Again, maximum likelihood estimation

$$\min_{\mathbf{W}} \hat{\mathbb{E}} \left[-\log \frac{\exp(\langle \mathsf{X}, \mathbf{w}_{\mathsf{Y}} \rangle)}{\sum_{l=1}^{c} \exp(\langle \mathsf{X}, \mathbf{w}_l \rangle)} \right]$$

- Prediction: $\hat{\mathsf{y}} = \operatorname{argmax}_k \Pr(\mathsf{Y} = k | \mathsf{X} = \mathbf{x}; \mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_c])$


Questions
?
?
Answers
?