CS480/680: Introduction to Machine Learning

Lec 03: Logistic Regression

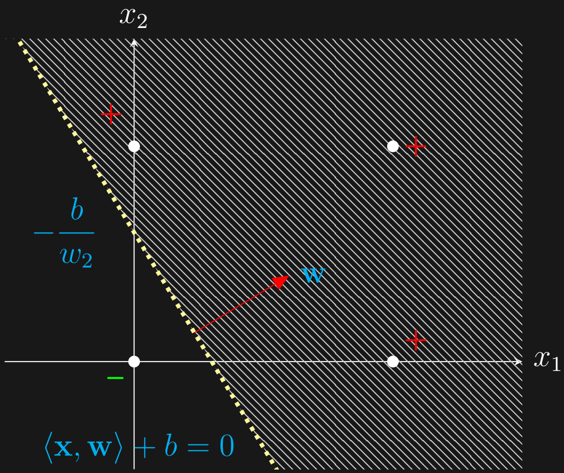
Yaoliang Yu

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS
DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE

Jan 16, 2025

Predicting with Confidence

- Recall that $\hat{y} = \mathrm{sign}(\langle \mathbf{x}, \mathbf{w} \rangle)$
- How confident are we about the prediction $\hat{y}$?
- Can use $|\langle \mathbf{x}, \mathbf{w} \rangle|$ as an indication
  - in fact was used in multi-class preceptron
  - real-valued: hard to interpret
  - many ways to transform into $[0, 1]$
- Better idea: learn confidence directly



Confidence Game

- $\mathsf{Y}_i \overset{i.i.d.}{\sim} \text{Bernoulli}(q)$ for some $q \in [0, 1]$
  - e.g., probability of snowing tomorrow
- How to evaluate a probabilistic forecast $\hat{p}$?
- Scoring function: $s : \mathcal{Y} \times [0, 1] \to \mathbb{R}$, $s(\mathsf{y}, p)$ scores the "fitness"
- Scoring rule: $\mathbb{S} : [0, 1] \times [0, 1] \to \mathbb{R}$, $\mathbb{S}(q, p) := \mathbb{E}_{\mathsf{Y} \sim \text{Bernoulli}(q)}[s(\mathsf{Y}, p)]$
- (Strict) properness (truthfulness): $q = \operatorname{argmin}_p \mathbb{S}(q, p)$
- Entropy: $\mathbb{H}(q) := \min_p \mathbb{S}(q, p)$, under properness, $\mathbb{H}(q) = \mathbb{S}(q, q)$

Logarithmic Loss

$$s(\mathsf{y}, p) := -\mathsf{y} \log p - (1 - \mathsf{y}) \log(1 - p)$$
$$\mathbb{S}(q, p) := -q \log p - (1 - q) \log(1 - p)$$
$$\mathbb{H}(q) := -q \log q - (1 - q) \log(1 - q)$$

- Indeed a proper scoring rule (could take ∞ value)
- The resulting entropy is exactly Shannon's entropy
- KL divergence: $\mathsf{KL}(q, p) := \mathbb{S}(q, p) - \mathbb{H}(q) \geq 0$, with equality iff $q = p$

I. J. Good. "Rational Decisions". *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 14, no. 1 (1952), pp. 107–114.

Introducing $\mathbb{X}$

$$
\begin{aligned}
s(\mathsf{y}, p(\mathbf{x})) &:= -\mathsf{y} \log p(\mathbf{x}) - (1-\mathsf{y}) \log(1-p(\mathbf{x})) \\
S(q(\mathbf{x}), p(\mathbf{x})) &:= -q(\mathbf{x}) \log p(\mathbf{x}) - (1-q(\mathbf{x})) \log(1-p(\mathbf{x})) \\
\mathbb{S}(q, p) &:= \mathbb{E}[-q(\mathsf{X}) \log p(\mathsf{X}) - (1-q(\mathsf{X})) \log(1-p(\mathsf{X}))]
\end{aligned}
$$

- $\mathsf{Y} \mid \mathsf{X} = \mathbf{x} \sim \mathrm{Bernoulli}(q(\mathbf{x}))$
- Observe that $S(q(\mathbf{x}), p(\mathbf{x})) = \mathbb{E}_{\mathsf{Y}|\mathsf{X}=\mathbf{x}} s(\mathsf{Y}, p(\mathbf{x}))$, $\mathbb{S}(q, p) := \mathbb{E}[s(\mathsf{Y}, p(\mathsf{X}))]$
- Parameterizing the probabilistic forecast, e.g. $p(\mathbf{x}; \mathbf{w}) = \mathrm{sgm}(\langle \mathbf{x}, \mathbf{w} \rangle)$
- Minimum score estimation:

$$
\min_{\mathbf{w}} \ \hat{\mathbb{E}}[s(\mathsf{Y}, p(\mathsf{X}; \mathbf{w}))]
$$

Max Conditional Likelihood

- Model postulates $\mathsf{Y}|\mathsf{X} = \mathbf{x} \sim \mathrm{Bernoulli}(p(\mathbf{x}; \mathbf{w}))$, i.e. $\Pr(\mathsf{Y} = 1|\mathsf{X} = \mathbf{x}) = p(\mathbf{x}; \mathbf{w})$
- Given $(\mathbf{X}_i, \mathsf{y}_i), i = 1, \ldots, n$, assume independence:

$$\begin{aligned}\Pr(\mathsf{Y}_1 = \mathsf{y}_1, \ldots, \mathsf{Y}_n = \mathsf{y}_n | \mathsf{X}_1 = \mathbf{x}_1, \ldots, \mathsf{X}_n = \mathbf{x}_n) &= \prod_{i=1}^{n} \Pr(\mathsf{Y}_i = \mathsf{y}_i | \mathsf{X}_i = \mathbf{x}_i) \\ &= \prod_{i=1}^{n} [p(\mathbf{x}_i; \mathbf{w})]^{\mathsf{y}_i} [1 - p(\mathbf{x}_i; \mathbf{w})]^{1-\mathsf{y}_i}\end{aligned}$$

- Maximizing the conditional log-likelihood:

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

Two Extremes

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

- What is the solution if $p(\mathbf{x}; \mathbf{w}) = p(\mathbf{w})$?
  - i.e. use the same confidence p for every data point
- What is the solution if $p(\mathbf{x}; \mathbf{w}) = p(\mathbf{x})$?
  - i.e. every data point uses its own confidence $p(\mathbf{x})$

The Logit Transform

- $p(\mathbf{x}; \mathbf{w}) : \mathcal{X} \to [0, 1]$, how to parameterize using $\mathbf{w}$?
  - $p(\mathbf{x}; \mathbf{w}) = \langle \mathbf{x}, \mathbf{w} \rangle$?
  - $\log p(\mathbf{x}; \mathbf{w}) = \langle \mathbf{x}, \mathbf{w} \rangle$?
- Logit transform: $\log \frac{p(\mathbf{x};\mathbf{w})}{1-p(\mathbf{x};\mathbf{w})} = \langle \mathbf{x}, \mathbf{w} \rangle$
  - i.e., the odds ratio is an affine function
- Equivalently, the sigmoid transformation: $p(\mathbf{x}; \mathbf{w}) = \mathrm{sgm}(\langle \mathbf{x}, \mathbf{w} \rangle) := \frac{1}{1+\exp(-\langle \mathbf{x},\mathbf{w} \rangle)}$

Logistic Regression

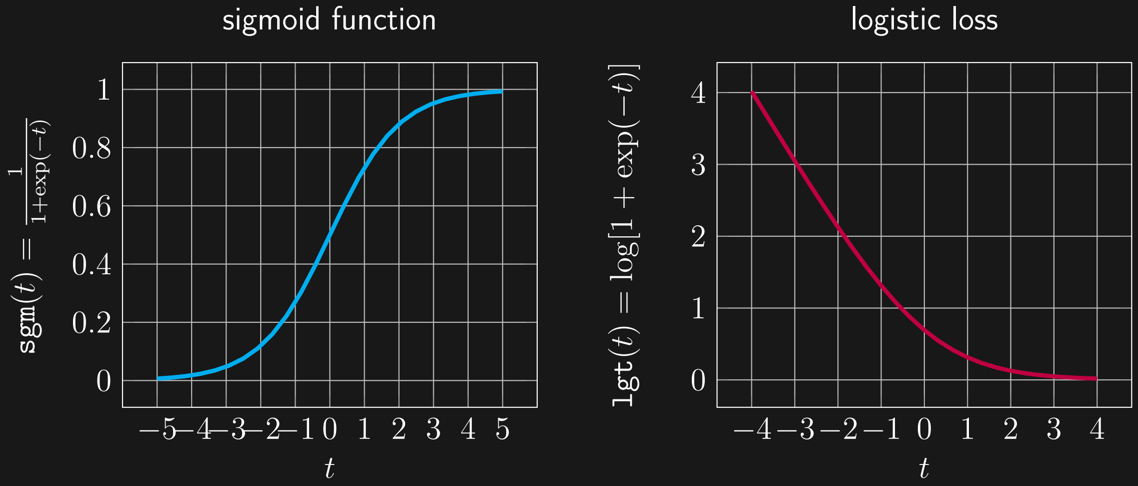
$$\min_{\mathbf{w}} \sum_{i=1}^{n} -\mathsf{y}_i \log[p(\mathbf{x}_i;\mathbf{w})] - (1-\mathsf{y}_i)\log[1-p(\mathbf{x}_i;\mathbf{w})]$$

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

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

- Note the label encoding $\mathsf{y}_i \in \{0,1\}$; if instead, $\mathsf{y}_i \in \{\pm 1\}$, then

$$\min_{\mathbf{w}} \sum_{i=1}^{n} \underbrace{\log\left[1+\exp\left(-\mathsf{y}_i\langle \mathbf{x}_i,\mathbf{w}\rangle\right)\right]}_{\text{logistic loss}}$$



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

Prediction

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

- $\hat{\mathrm{y}} = 1$ iff $p(\mathrm{x};\mathbf{w}) = \Pr(\mathsf{Y} = 1 | \mathsf{X} = \mathrm{x}) > \frac{1}{2}$ iff $\langle \mathbf{x}, \mathbf{w} \rangle > 0$
- Decision boundary remains to be $H := \{\mathrm{x} : \langle \mathbf{x}, \mathbf{w} \rangle = 0\}$
- Can predict $\hat{\mathrm{y}} = \mathrm{sign}(\langle \mathbf{x}, \mathbf{w} \rangle)$ as before, but now with confidence $p(\mathrm{x};\mathbf{w})$

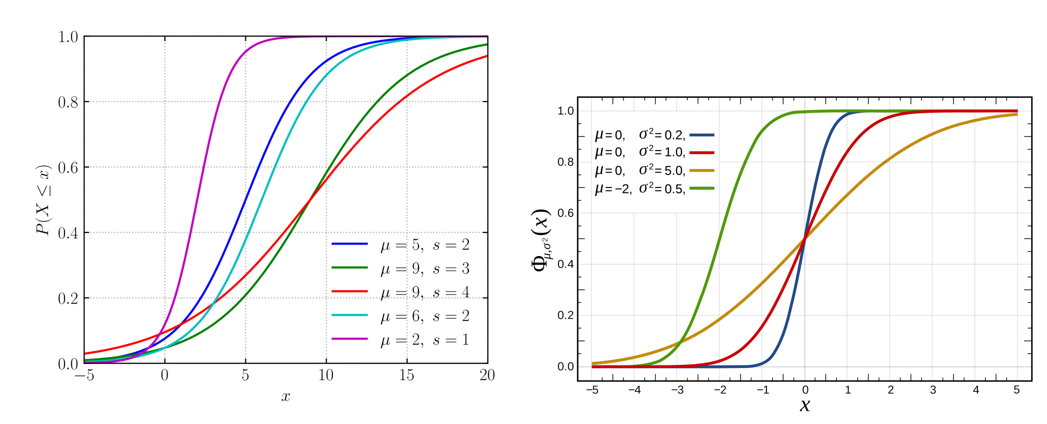
More than a Classification Algorithm

- Logistic regression estimates the posterior probability $\eta(\mathbf{x}) := \Pr(\mathsf{Y} = 1 | \mathsf{X} = \mathbf{x})$ under the linear odds ratio assumption
  - confidence is meaningless if the assumption is way off
- Classification itself only requires comparing $\eta(\mathbf{x})$ with $\frac{1}{2}$
- Possible to do the comparison without estimating $\eta(\mathbf{x})$ explicitly!
  - sufficient but not necessary, be lazy... SVM later

Beyond Logistic

$$p(\mathbf{x};\mathbf{w}) = F(\langle \mathbf{x}, \mathbf{w} \rangle)$$

- $F : \mathbb{R} \to [0,1]$, increasing: any cumulative distribution function (cdf) would do
- Logistic distribution: $F(x;\mu,s) = \frac{1}{1+\exp(-\frac{x-\mu}{s})}$
  - with mean μ and variance $s^2\pi^2/3$
  - sigmoid is exactly when $\mu = 0$ and $s = 1$


1.0
0.8
0.6
0.4
0.2
0.0
$P(X \le x)$
−5
0
5
10
15
20
x
$\mu = 5,\ s = 2$
$\mu = 9,\ s = 3$
$\mu = 9,\ s = 4$
$\mu = 6,\ s = 2$
$\mu = 2,\ s = 1$
$\Phi_{\mu,\sigma^2}(x)$
$\mu = 0,\quad \sigma^2 = 0.2,$
$\mu = 0,\quad \sigma^2 = 1.0,$
$\mu = 0,\quad \sigma^2 = 5.0,$
$\mu = -2,\quad \sigma^2 = 0.5,$
−5 −4 −3 −2 −1 0 1 2 3 4 5
x

Solving Logistic Regression

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

- Newton's algorithm:

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \cdot [\nabla^2 f(\mathbf{w})]^{-1} \cdot \nabla f(\mathbf{w})$$

- The gradient $\nabla f(\mathbf{w}) = \mathbf{X}(\hat{\mathbf{p}} - \frac{\mathbf{y}+1}{2})$: compare prediction $\hat{\mathbf{p}}$ with true label $\frac{\mathbf{y}+1}{2}$
- The Hessian $\nabla^2 f(\mathbf{w}) = \sum_i \hat{p}_i(1 - \hat{p}_i)\mathbf{x}_i\mathbf{x}_i^\top$: weighted by confidence
- The confidence $\hat{p}_i = \mathrm{sgm}(\langle \mathbf{x}_i, \mathbf{w} \rangle)$

Linear Regression vs. Logistic Regression

• least-squares: $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	• cross-entropy: $\sum_{i=1}^{n} -\frac{1+\mathsf{y}_i}{2}\log\hat{p}_i - \frac{1-\mathsf{y}_i}{2}\log(1-\hat{p}_i)$
• prediction: $\hat{y}_i = \langle \mathbf{x}_i, \mathbf{w}\rangle$	• prediction: $\hat{\mathsf{y}}_i = \mathrm{sign}(\langle \mathbf{x}_i, \mathbf{w}\rangle)$, $\hat{p}_i = \mathbf{sgm}(\langle \mathbf{x}_i, \mathbf{w}\rangle)$
• objective: $\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2$	• objective: $\mathsf{KL}(\frac{1+\mathbf{y}}{2}\|\hat{\mathbf{p}})$
• grad: $\mathbf{w} \leftarrow \mathbf{w} - \eta\mathbf{X}(\hat{\mathbf{y}} - \mathbf{y})$	• grad: $\mathbf{w} \leftarrow \mathbf{w} - \eta\mathbf{X}(\hat{\mathbf{p}} - \frac{1+\mathbf{y}}{2})$
• Newton: $\mathbf{w} \leftarrow \mathbf{w} - \eta(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}(\hat{\mathbf{y}} - \mathbf{y})$	• Newton: $\mathbf{w} \leftarrow \mathbf{w} - \eta(\mathbf{X}\hat{S}\mathbf{X}^\top)^{-1}\mathbf{X}(\hat{\mathbf{p}} - \frac{1+\mathbf{y}}{2})$

- Diagonal weight matrix $\hat{S} = \mathrm{diag}\left(\hat{\mathbf{p}} \odot (1 - \hat{\mathbf{p}})\right)$
- Logistic regression = iteratively weighted linear regression

More than 2 Classes

- Softmax parameterization:

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

  - nonnegative and sum to 1

- Encode $\mathsf{y} \in \{1, \ldots, c\}$
- Minimizing again the logarithmic loss:

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