CS489/698: Intro to ML

Lecture 04: Logistic Regression

UNIVERSITY OF WATERLOO

Outline

- Announcements
- Bernoulli model
- Logistic regression
- Computation

Outline

- Announcements
- Bernoulli model
- Logistic regression
- Computation

Announcements

- Assignment 1 due next Tuesday

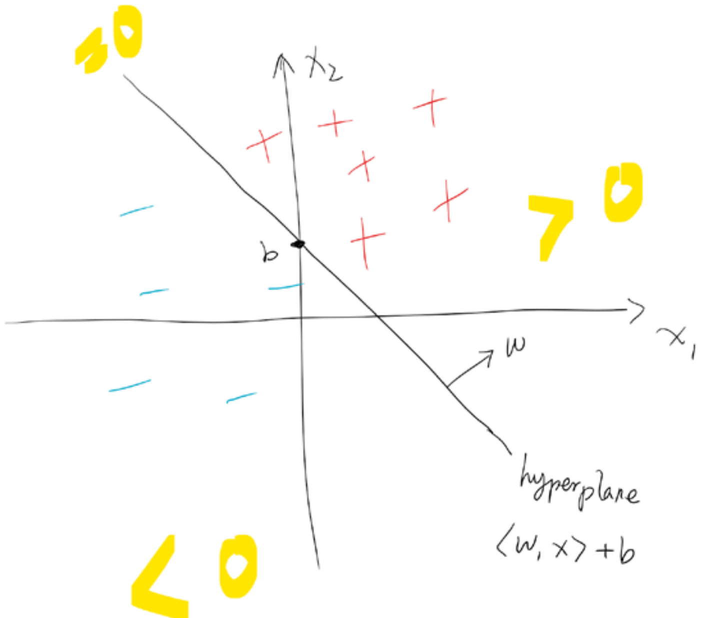
Outline

- Announcements
- Bernoulli model
- Logistic regression
- Computation

UNIVERSITY OF WATERLOO

Classification revisited

- $\hat{y} = \text{sign}(\ x^T w + b\ )$
- How confident we are about $\hat{y}$?
- $|x^T w + b|$ seems a good indicator
  - real-valued; hard to interpret
  - ways to transform into [0,1]
- Better(?) idea: learn confidence directly



Conditional probability

- $P(Y=1 \mid X=x)$: conditional on seeing x, what is the chance of this instance being positive, i.e., $Y=1$?
  - obviously, value in [0,1]

- $P(Y=0 \mid X=x) = 1 - P(Y=1 \mid X=x)$, if two classes
  - more generally, sum to 1

Notation (Simplex). $\Delta_{k-1} := \{ p \text{ in } R^k : p \geq 0, \Sigma_i p_i = 1 \}$

Reduction to a harder problem

- $P(Y=1 \mid X=x) = E(1_{Y=1} \mid X=x)$

$$1_A = \begin{cases} 1, & A \text{ is true} \\ 0, & A \text{ is false} \end{cases}$$

- Let $Z = 1_{Y=1}$, then regression function for (X, Z)
  - use linear regression for binary Z?
- Exploit structure!
  - conditional probabilities are in a simplex
- Never reduce to unnecessarily harder problem

Bernoulli model

- Let P(Y=1 | X=x) = p(x; w), parameterized by w

- Conditional likelihood on {(x_1, y_1), … (x_n, y_n)}:

$$\mathbf{P}(Y_1 = y_1, \ldots, Y_n = y_n | X_1 = \mathbf{x_1}, \ldots, X_n = \mathbf{x}_n)$$

  - simplifies if independence holds

$$\prod_{i=1}^{n} \mathbf{P}(Y_i = y_i | X_i = \mathbf{x}_i) = \prod_{i=1}^{n} p(\mathbf{x}_i; \mathbf{w})^{y_i} (1 - p(\mathbf{x}_i; \mathbf{w}))^{1-y_i}$$

  - Assuming y_i is {0,1}-valued

UNIVERSITY OF WATERLOO

Naïve solution

$$\prod_{i=1}^{n} p(\mathbf{x}_i; \mathbf{w})^{y_i} (1 - p(\mathbf{x}_i; \mathbf{w}))^{1-y_i}$$

- Find w to maximize conditional likelihood
- What is the solution if p(x; w) does not depend on x?
- What is the solution if p(x; w) does not depend on ?

Generalized linear models (GLM)

- $y \sim \text{Bernoulli}(p)$; $p = p(x; w)$ natural parameter
  - Logistic regression

- $y \sim \text{Normal}(\boldsymbol{\mu}, \sigma^2)$; $\boldsymbol{\mu} = \boldsymbol{\mu}(x; w)$
  - (weighted) least-squares regression

- GLM: $y \sim \exp(\ \theta\ \varphi(y) - A(\theta)\ )$

log-partition function

sufficient statistics

Outline

- Announcements
- Bernoulli model
- Logistic regression
- Computation

UNIVERSITY OF WATERLOO

Logit transform

- p(x; w) = $w^T x$? p >=0 not guaranteed…
- log p(x; w) = $w^T x$? better!
  - LHS negative, RHS real-valued…
- Logit transform $$\log \frac{p(\mathbf{x}; \mathbf{w})}{1 - p(\mathbf{x}; \mathbf{w})} = \mathbf{w}^\top \mathbf{x}$$ odds ratio
- Or equivalently $$p(\mathbf{x}; \mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{x})}$$

Prediction with confidence

$$p(\mathbf{x}; \mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{x})}$$

- $\hat{y} = 1$ if $p = P(Y=1 \mid X=x) > ½$ iff $w^\top x > 0$
- Decision boundary $w^\top x = 0$
- $\hat{y} = \text{sign}(w^\top x)$ as before, but with confidence $p(x; w)$

Not just a classification algorithm

- Logistic regression does more than classification
  - it estimates conditional probabilities
  - under the logit transform assumption

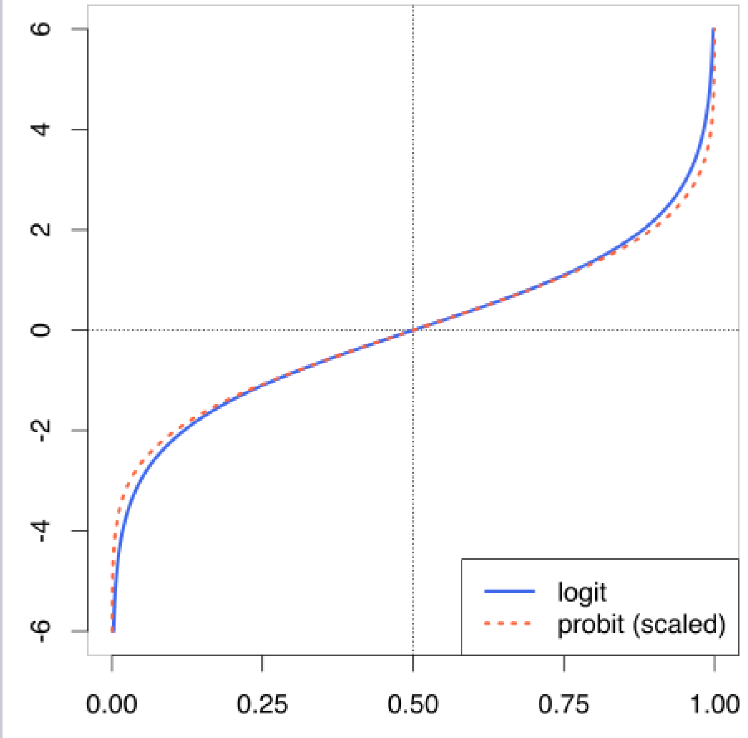
- Having confidence in prediction is nice
  - the price is an assumption that may or may not hold

- If classification is the sole goal, then doing extra work
  - as shall see, SVM only estimates decision boundary

UNIVERSITY OF WATERLOO

More than logistic regression

- F(p) transforms p from [0,1] to R
- Then, equating F(p) to a linear function $w^T x$
- But, there are many other choices for F!
    - precisely the inverse of any distribution function!



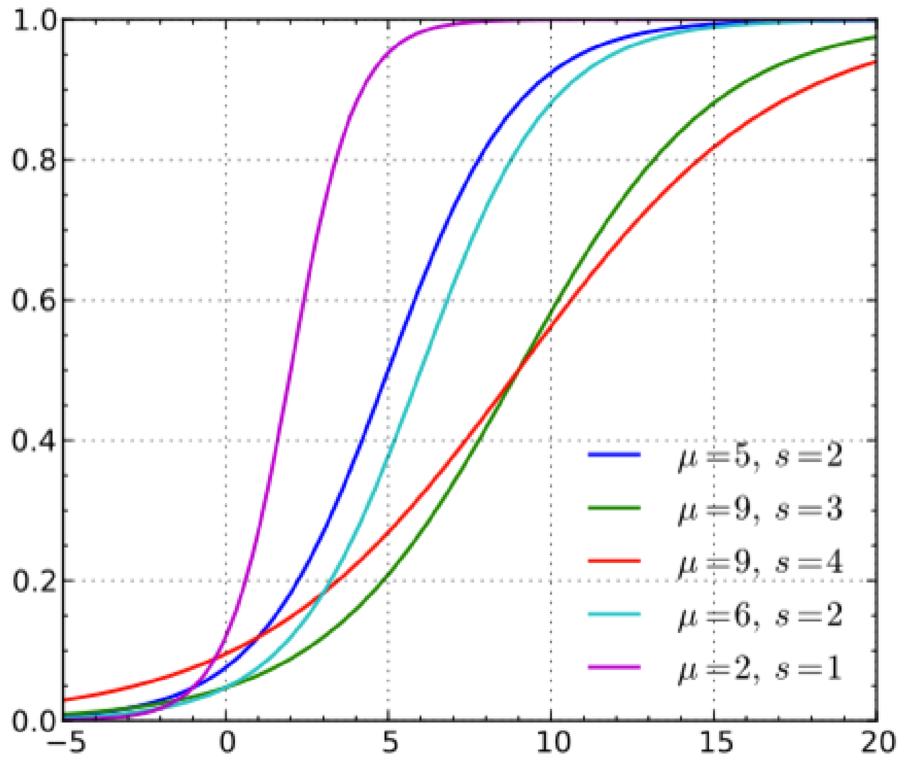
Comparison of the logit function with a scaled probit (i.e. the inverse CDF of the normal distribution), comparing $\operatorname{logit}(x)$ vs. $\Phi^{-1}(x)/\sqrt{\frac{\pi}{8}}$, which makes the slopes the same at the origin.

Logistic distribution

- Cumulative Distribution Function

$$F(x;\mu,s) = \frac{1}{1+\exp(-\frac{x-\mu}{s})}$$

- Mean mu, variance $s^2\pi^2/3$



Outline

- Announcements
- Bernoulli model
- Logistic regression
- Computation

Maximum likelihood

$$\prod_{i=1}^{n} p(\mathbf{x}_i;\mathbf{w})^{y_i}(1-p(\mathbf{x}_i;\mathbf{w}))^{1-y_i}$$

$$p(\mathbf{x};\mathbf{w}) = \frac{1}{1+\exp(-\mathbf{w}^\top\mathbf{x})}$$

- Minimize negative log-likelihood

$$\sum_i \log(e^{(1-y_i)\mathbf{w}^\top\mathbf{x}_i} + e^{-y_i\mathbf{w}^\top\mathbf{x}_i}) \equiv \sum_i \log(1+e^{-\tilde{y}_i\mathbf{w}^\top\mathbf{x}_i})$$

Newton's algorithm

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t [\nabla^2 f(\mathbf{w}_t)]^{-1} \cdot \nabla f(\mathbf{w}_t)$$

$$\nabla f(\mathbf{w}_t) = X^\top (\mathbf{p} - \mathbf{y})$$

$$\boxed{\nabla^2 f(\mathbf{w}_t) = \sum_i p_i(1 - p_i)\mathbf{x}_i \mathbf{x}_i^\top}$$

PSD

$$p_i = \frac{1}{1 + e^{-\mathbf{w}_t^\top \mathbf{x}_i}}$$

Uncertain predictions get bigger weight

- η = 1: iterative weighted least-squares

UNIVERSITY OF WATERLOO

A word about implementation

- Numerically computing exponential can be tricky
  - easily underflows or overflows
- The usual trick
  - estimate the range of the exponents
  - shift the mean of the exponents to 0

Robustness

$$\ell(t) = \log(1 + e^t) \quad L(\hat{y}, y) = \ell(-\hat{y}y) \quad \hat{y} = \mathbf{w}^\top \mathbf{x}$$

- Bounded derivative

$$\ell'(t) = \frac{e^t}{1 + e^t} = \frac{1}{1 + e^{-t}}$$

- Variational exponential

Larger exp loss gets smaller weights

$$\log(1 + e^t) = \min_{0 \le \eta \le 1} \boxed{\eta e^t} - \log(\eta) + \eta - 1$$

More than 2 classes

- Softmax

$$\mathbf{P}(Y = c | \mathbf{x}, W) = \frac{\exp(\mathbf{w}_c^\top \mathbf{x})}{\sum_{q=1}^{k} \exp(\mathbf{w}_q^\top \mathbf{x})}$$

- Again, nonnegative and sum to 1

- Negative log-likelihood (y is one-hot)

$$-\log \prod_{i=1}^{n} \prod_{c=1}^{k} p_{ic}^{y_{ic}} = -\sum_{i=1}^{n} \sum_{c=1}^{k} y_{ic} \log p_{ic}$$

Questions?