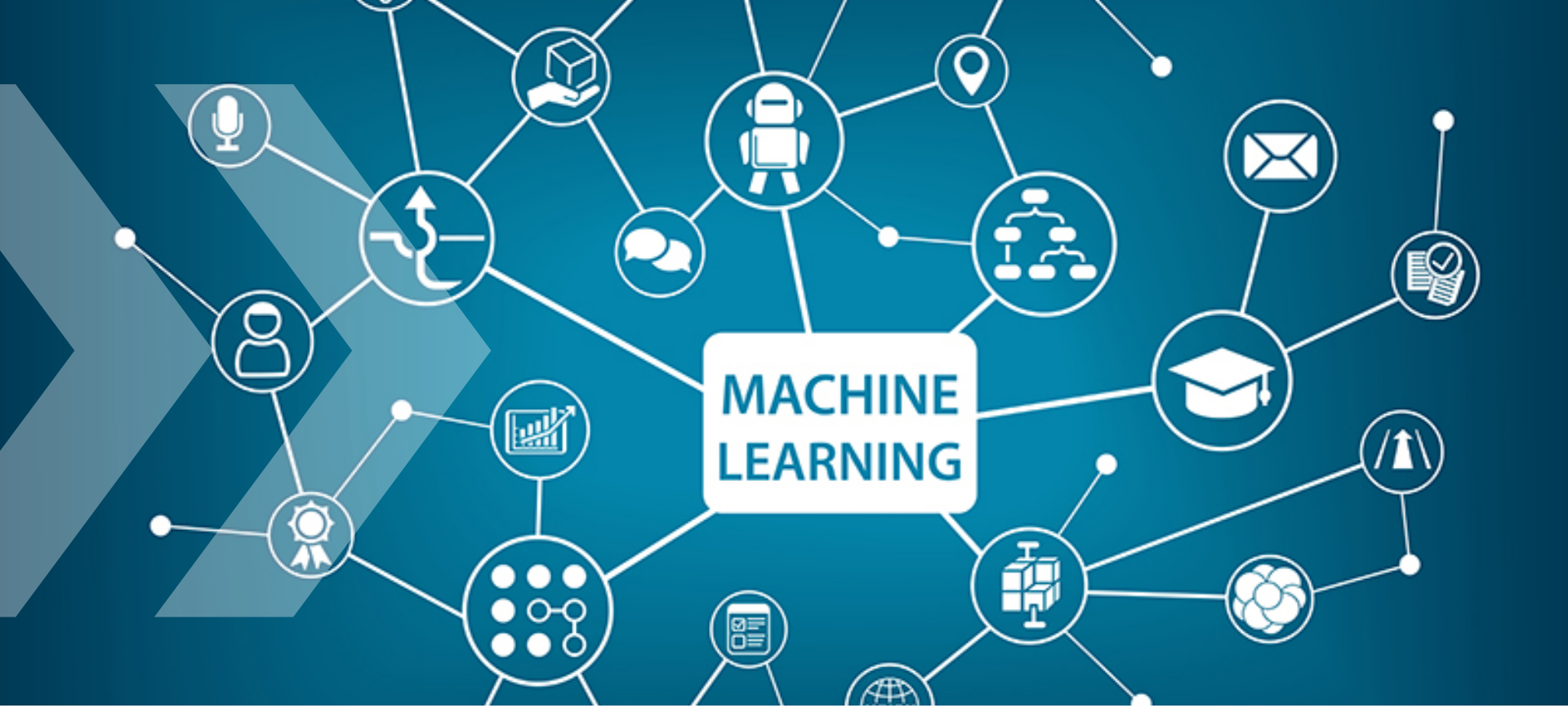


CS480/680: Intro to ML

Lecture 06: Logistic Regression

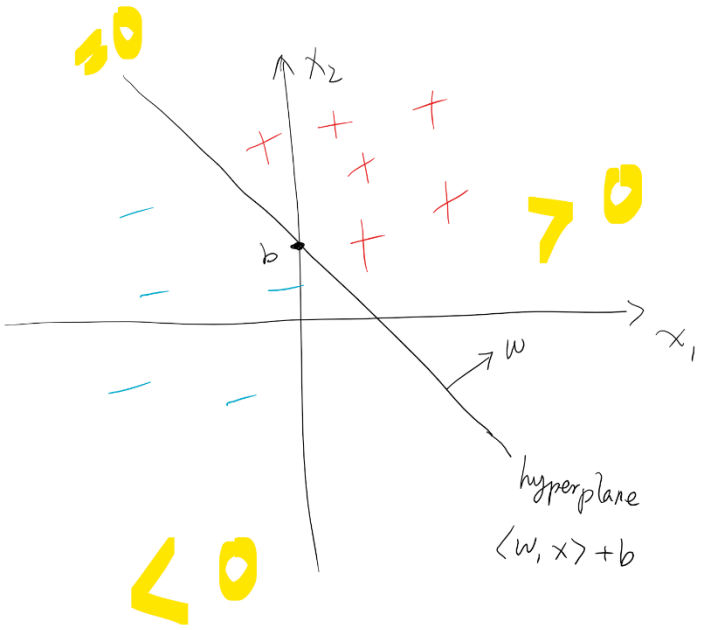
UNIVERSITY OF WATERLOO

Outline

- Bernoulli model
- Logistic regression
- Computation

Classification revisited

- $\hat{y} = \text{sign}(\mathbf{x}^T\mathbf{w} + b)$
- How confident we are about $\hat{y}$?
- $|\mathbf{x}^T\mathbf{w} + b|$ 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 | 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 | X=**x**) = 1 – P(Y=1 | X=**x**), if two classes
  - more generally, sum to 1

Notation (Simplex). $\Delta_{c-1} := \{ \mathbf{p} \text{ in } R^c : \mathbf{p} \geq 0, \Sigma_k p_k = 1 \}$

Bernoulli model

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

- Conditional likelihood on {($\mathbf{x}_1$, y_1), … ($\mathbf{x}_n$, 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

Naïve solution

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

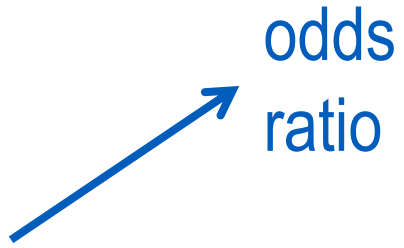
- Find **w** to maximize conditional likelihood
- What is the solution if p(**x**; **w**) does not depend on **x**?
- What is the solution if p(**x**; **w**) does not depend on **w**?

Outline

- Announcements
- Bernoulli model
- Logistic regression
- Computation

Logit transform

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



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

Prediction with confidence

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

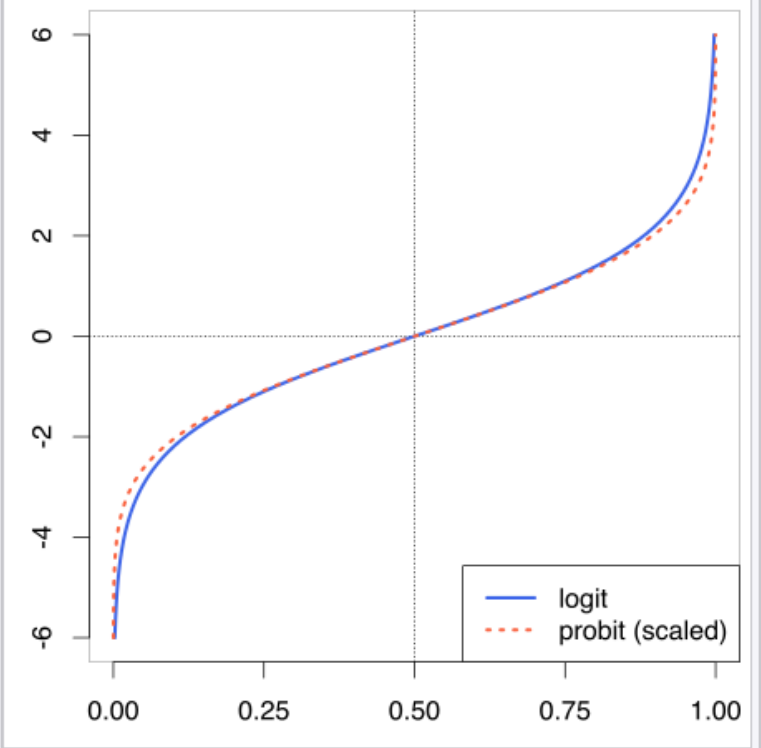
- $\hat{y} = 1$ if $p = P(Y=1 \mid X=x) > ½$ iff $\mathbf{w}^T\mathbf{x} > 0$
- Decision boundary $\mathbf{w}^T\mathbf{x} = 0$
- $\hat{y} = \text{sign}(\mathbf{w}^T\mathbf{x})$ as before, but with confidence $p(x; w)$

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

More than logistic regression

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



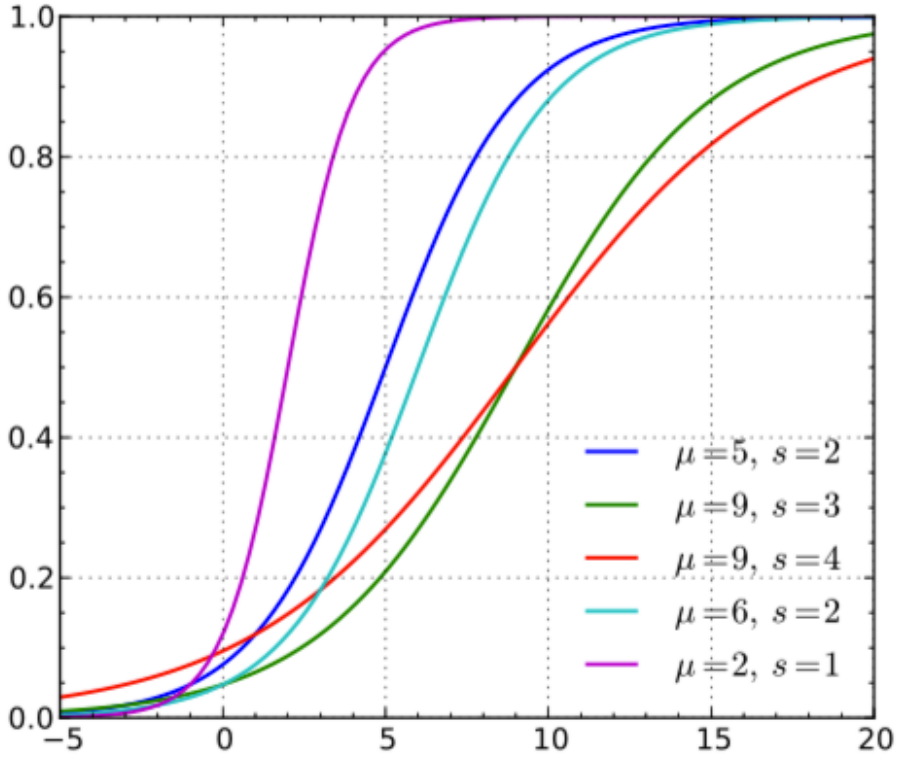
Comparison of the logit function with a scaled probit (i.e. the inverse CDF of the normal distribution), comparing $\operatorname{logit}(x)$ vs. $\Phi^{-1}(x)/\sqrt{\frac{\pi}{8}}$, which makes the slopes the same at the origin.

Logistic distribution

- Cumulative Distribution Function

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

- Mean mu, variance $s^2\pi^2/3$
- Sigmoid: mu=0, s = 1



Outline

- Announcements
- Bernoulli model
- Logistic regression
- Computation

Maximum likelihood

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

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

cross-entropy $$-\frac{1}{n} \sum_{i=1}^{n} y_i \log p_i + (1 - y_i) \log(1 - p_i)$$

- Minimize negative log-likelihood

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

objective function *f*

Newton's algorithm

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

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

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

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

Comparison

$$\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

$$\hat{y}_i = \mathbf{w}^\top \mathbf{x}_i$$

$$\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2$$

$$\mathbf{w}- = \eta X(\hat{\mathbf{y}} - \mathbf{y})$$

$$\mathbf{w}- = \eta (XX^\top)^{-1} X(\hat{\mathbf{y}} - \mathbf{y})$$

$$\sum_{i=1}^{n} y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)$$

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

$$\mathsf{KL}(\mathbf{y} \| \hat{\mathbf{p}})$$

$$\mathbf{w}- = \eta X(\hat{\mathbf{p}} - \mathbf{y})$$

$$\mathbf{w}- = \eta (XSX^\top)^{-1} X(\hat{\mathbf{p}} - \mathbf{y})$$

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

More than 2 classes

- Softmax

$$\mathbf{P}(Y = k|\mathbf{x}, W) = \frac{\exp(\mathbf{w}_k^\top \mathbf{x})}{\sum_{l=1}^{c} \exp(\mathbf{w}_l^\top \mathbf{x})}$$

- Again, nonnegative and sum to 1

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

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

Questions?