CS 4758: Logistic Regression

Ashutosh Saxena

Cornell University

CS 4758 announcements

- HW1 due in class this Thursday.
- Project proposal due Feb 15 (or earlier).
  - See template on the webpage.
- Project sprint 1 report/presentation on Mar 4.

Lecture overview

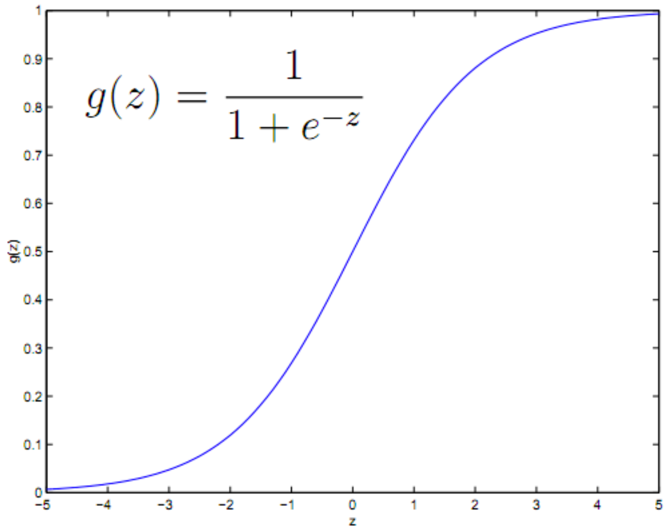
- Basics: Robot Kinematics.
- Algorithms:
  - Gradient descent (different variants).
  - Newton (today)
- Learning algorithms
  - K-NN
  - Supervised learning setting
  - Training/testing/cross-validation data-set. Overfitting. Importance of data-set.
  - Linear regression
  - Logistic Regression (today)
  - 3D Features (Feb 9)
- Software
  - ROS
  - PCL (Feb 14)

- Markov Chains, MDP, reinforcement learning. (Feb 16 onwards)

Classification

- Y = {0,1}
- E.g., spam vs non-spam
- Chair vs no-chair.

- Pickable object vs not.

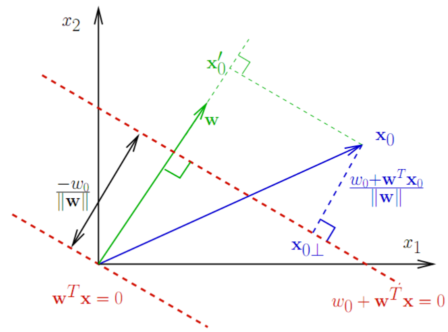


- Values of θ (or w) change the location of transition and its sharpness.

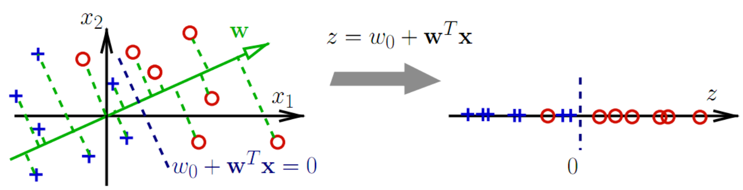
$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}},$$

Review: linear classification

- Linear projections



- Linear classification ⇔ 1-D dimensionality reduction



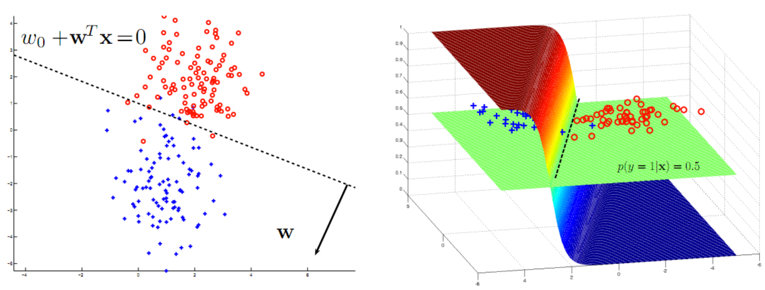
Review: logistic regression model

- Binary classification, $\mathcal{Y} = \{0, 1\}$
- Model the posterior

$$p(y = 1 \mid \mathbf{x}) = g(w_0 + \mathbf{w}^T\mathbf{x}) = \frac{1}{1 + \exp(-w_0 - \mathbf{w}^T\mathbf{x})}$$

- Linear decision boundary:

$$\hat{y} = 1 \Leftrightarrow g(w_0 + \mathbf{w}^T\mathbf{x}) > \tfrac{1}{2} \Leftrightarrow w_0 + \mathbf{w}^T\mathbf{x} = 0$$



Likelihood under the logistic model

- Regression: observe values, measure residuals under the model.
- Logistic regression: observe labels, measure their probability under the model.

$$p\left(y_i \mid \mathbf{x}_i; \mathbf{w}\right) = \begin{cases} g\left(w_0 + \mathbf{w}^T \mathbf{x}_i\right) & \text{if } y_i = 1, \\ 1 - g\left(w_0 + \mathbf{w}^T \mathbf{x}_i\right) & \text{if } y_i = 0 \end{cases}$$

Likelihood under the logistic model

- Regression: observe values, measure residuals under the model.
- Logistic regression: observe labels, measure their probability under the model.

$$\begin{aligned} p\left(y_i \mid \mathbf{x}_i; \mathbf{w}\right) &= \begin{cases} g\left(w_0 + \mathbf{w}^T \mathbf{x}_i\right) & \text{if } y_i = 1, \\ 1 - g\left(w_0 + \mathbf{w}^T \mathbf{x}_i\right) & \text{if } y_i = 0 \end{cases} \\ &= g\left(w_0 + \mathbf{w}^T \mathbf{x}_i\right)^{y_i} \left(1 - g\left(w_0 + \mathbf{w}^T \mathbf{x}_i\right)\right)^{1-y_i}. \end{aligned}$$

Likelihood under the logistic model

- Regression: observe values, measure residuals under the model.
- Logistic regression: observe labels, measure their probability under the model.

$$p(y_i \mid \mathbf{x}_i; \mathbf{w}) = \begin{cases} g(w_0 + \mathbf{w}^T\mathbf{x}_i) & \text{if } y_i = 1, \\ 1 - g(w_0 + \mathbf{w}^T\mathbf{x}_i) & \text{if } y_i = 0 \end{cases}$$
$$= g(w_0 + \mathbf{w}^T\mathbf{x}_i)^{y_i} \left(1 - g(w_0 + \mathbf{w}^T\mathbf{x}_i)\right)^{1-y_i}.$$

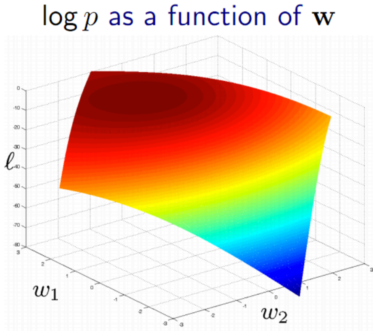
- The log-likelihood of $\mathbf{w}$:

$$\log p(Y|X; \mathbf{w}) = \sum_{i=1}^{N} \log p(y_i \mid \mathbf{x}_i; \mathbf{w})$$
$$= \sum_{i=1}^{N} y_i \log g(w_0 + \mathbf{w}^T\mathbf{x}_i) + (1 - y_i) \log\left(1 - g(w_0 + \mathbf{w}^T\mathbf{x}_i)\right)$$

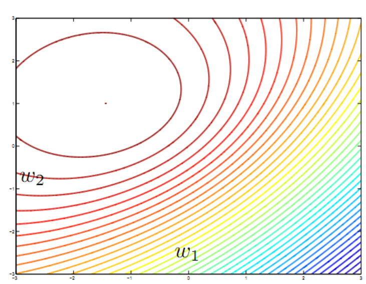
Visualizing the log-likelihood surface

- We will look at a 2D example, and assume $w_0 = 0$, i.e. our model will be $\hat{p}(y = 1|\mathbf{x}) = \sigma(w_1x_1 + w_2x_2)$.

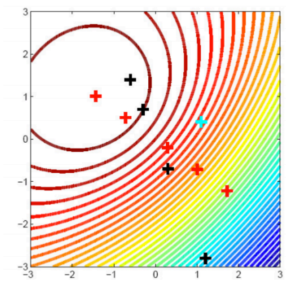
$\log p$ as a function of $\mathbf{w}$

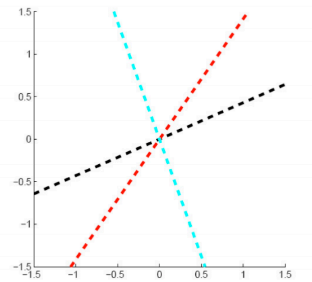


Contour plot: high/low

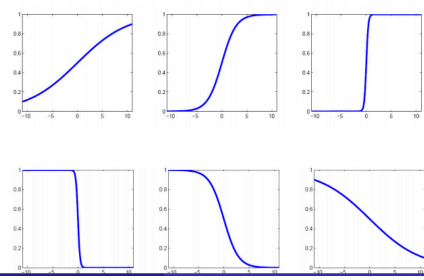


Mapping from boundaries to $\mathbf{w}$





- A line $\alpha\mathbf{w}$ in the w_1, w_2 space corresponds to a set of parralel decision boundaries of the form $\alpha\mathbf{w}^T\mathbf{x} = 0$.
- The sign of α determines the direction.



Derivation of g'(z)

(On blackboard.)

Update rule

$$w_j := w_j + \alpha \left(y^{(i)} - h_w(x^{(i)})\right) x_j^{(i)}$$

Generalized additive models

- As with regression we can extend this framework to arbitrary features (basis functions):

$$p(y = 1 \mid \mathbf{x}) \; = \; g\left(w_0 + \phi_1(\mathbf{x}) + \ldots + \phi_m(\mathbf{x})\right).$$

- Example: quadratic logistic regression in 2D

$$p(y = 1 \mid \mathbf{x}) \; = \; g\left(w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2\right).$$

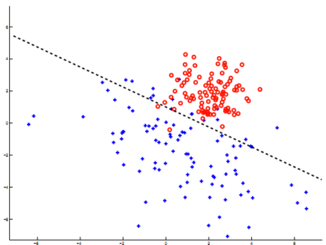
  - Decision boundary of this classifier:

$$w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2 = 0,$$

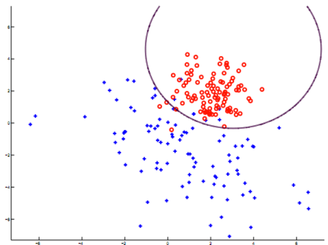
  i.e. it's a quadratic decision boundary.

Logistic regression: 2D example

Linear

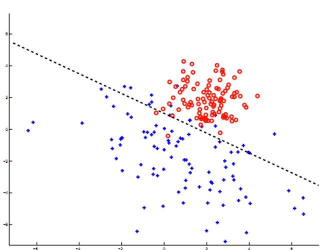


Quadratic

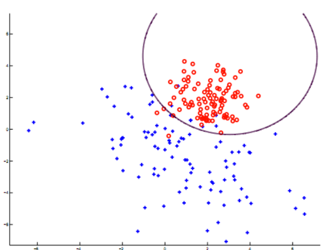


Logistic regression: 2D example

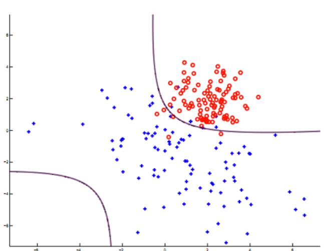
Linear



Quadratic



We can also include $x_1 x_2$:



Overfitting with logistic regression

- We can get the same decision boundary with an infinite number of settings for $\mathbf{w}$.
- When the data are *separable* by $w_0 + \alpha \mathbf{w}^T \mathbf{x} = 0$, what's the best choice for α?

$$p(y = 1 \mid \mathbf{x}) = \sigma(w_0 + \alpha \mathbf{w}^T \mathbf{x}).$$

- With $\alpha \to \infty$, we have $p(y_i|\mathbf{x}; w_0, \alpha\mathbf{w}) \to 1$.
- With $\alpha = \infty$ there is a continuum of w_0 that reach perfect separation.
- When the data are not separable, similar effect is present but more subtle.

MAP for logistic regression

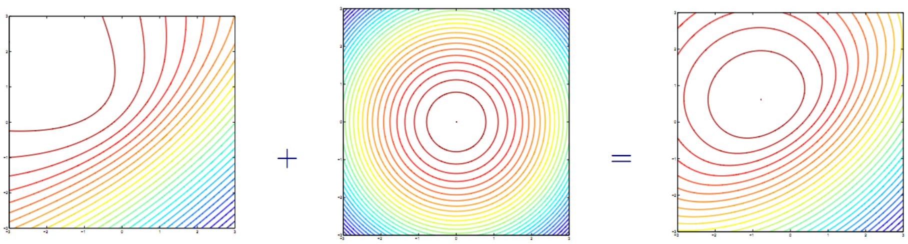
- Instead of $\log p(Y|X;\mathbf{w})$ the objective function (under the Gaussian prior) becomes:

$$\begin{aligned} \log \tilde{p}(Y|X, \mathbf{w}; \sigma) &= \log p(Y|X, \mathbf{w}) + \log p(\mathbf{w}; \sigma) \\ &= \sum_{i=1}^{N} \log p\left(y_i \mid \mathbf{x}_i; \mathbf{w}\right) - \frac{1}{2\sigma^2}(w_1^2 + w_2^2) + \text{const}(\mathbf{w}). \end{aligned}$$

- This is a *penalized log-likelihood* (or *log-posterior*).
- Note that $w_1^2 + w_2^2 = \|\mathbf{w}\|^2$.
- Setting σ^2 will affect the penalty we impose for a particular value of $\|\mathbf{w}\|$.

Penalized likelihood surface

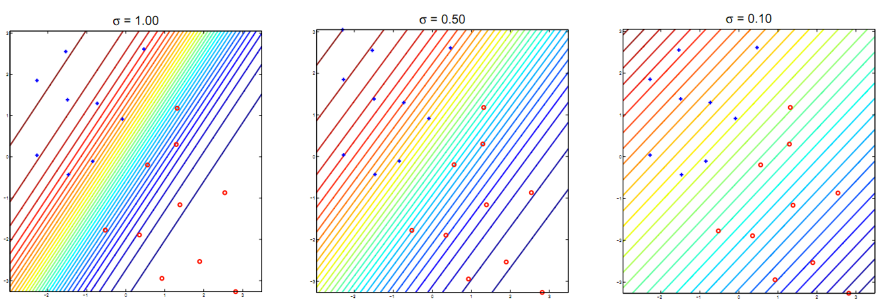
$\log p(Y|X;\mathbf{w})$ $\log p(\mathbf{w};\sigma)$ $\log \tilde{p}(Y|X,\mathbf{w};\sigma)$



- This is our objective function, and we can find its peak by gradient descent as before.
  - Need to modify the calculation of gradient and Hessian.

The effect of regularization: separable data

$$\log \tilde{p}(Y|X,\mathbf{w};\sigma) = \sum_{i=1}^{N} \log p\left(y_i \,|\, \mathbf{x}_i;\mathbf{w}\right) - \frac{1}{2\sigma^2}\|\mathbf{w}\|$$

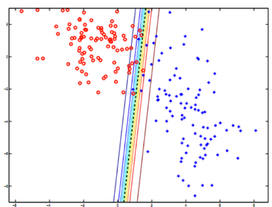


$\sigma^2 = 1$ $\sigma^2 = 0.5$ $\sigma^2 = 0.1$

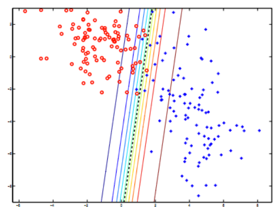
The effect of regularization

$$\log \tilde{p}(Y|X; \mathbf{w}, \sigma) = \sum_{i=1}^{N} \log p\left(y_i \,|\, \mathbf{x}_i; \mathbf{w}\right) - \frac{1}{2\sigma^2}||\mathbf{w}||$$

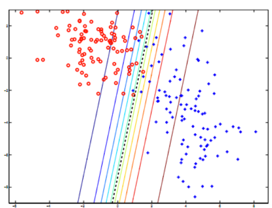
ML



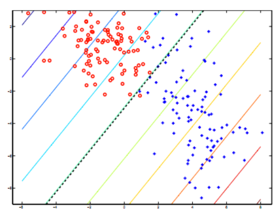
$\sigma^2 = 1$



$\sigma^2 = 0.1$



$\sigma^2 = 0.01$



How to optimize the objective function?

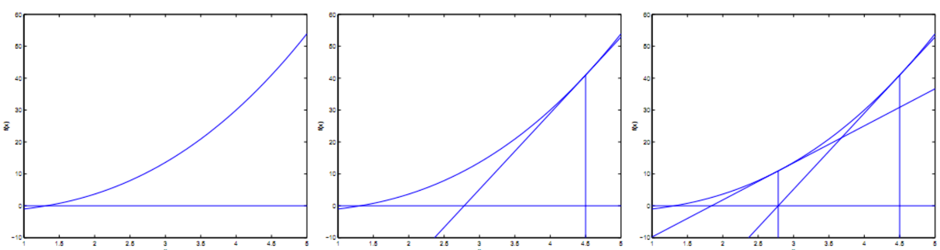
- Gradient descent
  - Coordinate descent
  - Stochastic gradient descent
  - Batch gradient descent

- Newton's method.

Newton's method

- Find zero of a function f(θ) = 0

$$\theta := \theta - \frac{f(\theta)}{f'(\theta)}.$$



Newton's method

- Maximize some function l(θ)
- f(θ) = l' (θ)

$$\theta := \theta - \frac{\ell'(\theta)}{\ell''(\theta)}.$$

$$\theta := \theta - H^{-1}\nabla_\theta \ell(\theta).$$

$$H_{ij} = \frac{\partial^2 \ell(\theta)}{\partial\theta_i \partial\theta_j}.$$

Softmax regression

- Y = {1,2,…,K}

- E.g., objects: {chair, table, monitor, none}.
- E.g., activities: {cooking, drinking, eating, none}.

Softmax idea

- Logistic regression, y={0,1}. h(x) was scalar.

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}},$$

$$p(y_i \mid \mathbf{x}_i; \mathbf{w}) = \begin{cases} \mathrm{g}(w_0 + \mathbf{w}^T \mathbf{x}_i) & \text{if } y_i = 1, \\ 1 - \mathrm{g}(w_0 + \mathbf{w}^T \mathbf{x}_i) & \text{if } y_i = 0 \end{cases}$$

- Now, it would be a vector. $\theta_1, \ldots, \theta_{k-1} \in \mathbb{R}^{n+1}$

$$\begin{aligned} p(y = i \mid x; \theta) &= \phi_i \\ &= \frac{e^{\eta_i}}{\sum_{j=1}^k e^{\eta_j}} \\ &= \frac{e^{\theta_i^T x}}{\sum_{j=1}^k e^{\theta_j^T x}} \end{aligned}$$

Softmax details (optional)

$$\theta_1, \ldots, \theta_{k-1} \in \mathbb{R}^{n+1}$$

$$\begin{aligned} p(y=i|x;\theta) &= \phi_i \\ &= \frac{e^{\eta_i}}{\sum_{j=1}^k e^{\eta_j}} \\ &= \frac{e^{\theta_i^T x}}{\sum_{j=1}^k e^{\theta_j^T x}} \end{aligned}$$

$$h_\theta(x) = \begin{bmatrix} \frac{\exp(\theta_1^T x)}{\sum_{j=1}^k \exp(\theta_j^T x)} \\ \frac{\exp(\theta_2^T x)}{\sum_{j=1}^k \exp(\theta_j^T x)} \\ \vdots \\ \frac{\exp(\theta_{k-1}^T x)}{\sum_{j=1}^k \exp(\theta_j^T x)} \end{bmatrix}$$

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^m \log p(y^{(i)}|x^{(i)};\theta) \\ &= \sum_{i=1}^m \log \prod_{l=1}^k \left(\frac{e^{\theta_l^T x^{(i)}}}{\sum_{j=1}^k e^{\theta_j^T x^{(i)}}} \right)^{1\{y^{(i)}=l\}} \end{aligned}$$

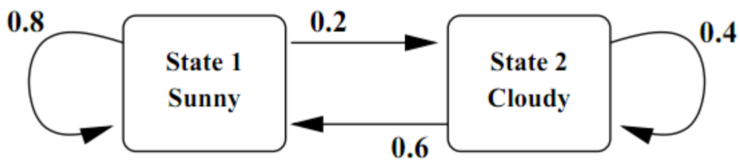
break

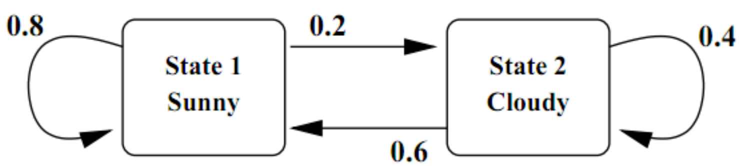
Markov Chains

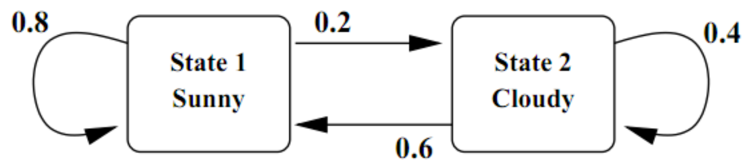
- Vector of probabilities at each step $\vec{\mathrm{p}}_n = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_r \end{bmatrix}$
- Transition probability

$$p_{ij} = \text{Prob(State } n+1 \text{ is } S_i \mid \text{State } n \text{ is } S_j$$

- Transition probability matrix $P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1r} \\ p_{21} & p_{22} & \cdots & p_{2r} \\ \vdots & & \ddots & \\ p_{r1} & & & p_{rr} \end{bmatrix}$
- Go from one step to next: $\vec{\mathrm{p}}_{n+1} = P\vec{\mathrm{p}}_n.$







$$P = \begin{bmatrix} 0.8 & 0.6 \\ 0.2 & 0.4 \end{bmatrix}$$

- That's all.