TIM 245 Lecture 10 (5/8/17)

Agenda

1) Review Homework 2

2) Review Lecture 9

3) Finish Logistic Regression

4) Gradient Descent

5) Extending binary classification to multiple classes

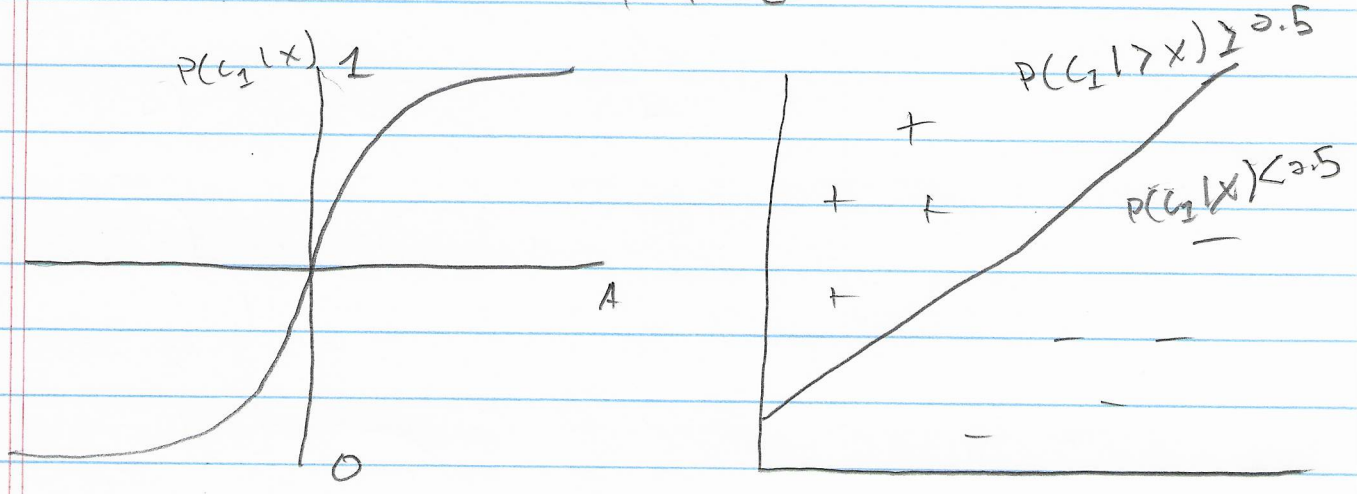
6) Introduction to Support Vector Machines

② Logistic Regression

Model the probability $P(C_1 | X)$ as a logistic function of X

$$P(C_1 | X) = \text{logit}^{-1}(\beta_0 + \beta^T X)$$

$$= \frac{1}{1 + e^{-(\beta_0 + \beta^T X)}}$$



Creates a linear decision boundry between the two classes : C_1, C_2

How do we find the coefficients that best Seperate the training data.

Let

$$\theta \triangleq \{\beta_0, \beta_1, \beta_2 \ldots \beta_m\}$$

$L(\theta \mid x, y) \triangleq$ likelihood of θ given training data (x, y)

$P(y \mid \theta, x) \triangleq$ probability of y given θ and x

$$y = \begin{cases} 1 & \text{when } C_1 \\ 0 & \text{when } C_2 \end{cases}$$

$$L(\theta \mid x, y) = P(y \mid \theta, x)$$

$$= P(y_1 \mid \theta, x_1) \ldots P(x_n \mid \theta, x_n)$$

$$= \prod_{i=1}^{n} P(y_i \mid \theta, x_i)$$

$$= \prod_{i=1}^{n} P(C_1 \mid \theta, x_i)^{y_i} \cdot P(C_2 \mid \theta, x_i)^{1-y_i}$$

$$P(C_1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + \beta^T x)}}$$

$$1 - P(C_1 \mid x)$$

We need to find θ_{MLE}

$$\theta_{MLE} = \arg\max_{\theta} \prod_{i=1}^{n} P(y_i | \theta, X)$$

No closed form solution. Need to solve using gradient descent.

(4) Gradient Descent

How do we optimize complex functions where there is no closed form solutions?

Let

$f(\theta) \triangleq$ cost function that we want to minimize, e.g. RSS

$\theta \triangleq$ function parameters, e.g β

Iterative Process:

1) Initialize the parameters θ to 0 or a random value

2) Compute the gradient, i.e. slope of the function so we know which direction to move

$$\Delta = \nabla_\theta f(\theta)$$

3) Update the parameter values in order to move in the direction of minimum cost

$$\theta_{i+1} = \theta_1 - \alpha \Delta$$

where α is a smoothing parameter that controls how far to move (step size)

4) Stop when we hit either a maximum number of iterations or epochs or Δ becomes small

In regular gradient descent, the gradient for each epoch is computed using the entire training dataset.

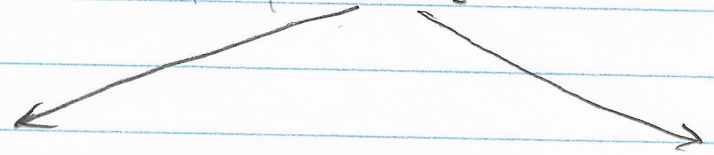
In Stochastic gradient descent the gradient is computed using the next instance in the training dataset (Batch size = 1)

Important to randomize the training dataset before applying Stochastic gradient descent.

⑤ Extending Binary Classification to multiple classes

Logistic Regression is a binary classifier, i.e. it finds the seperating line between two classes

Two approaches for decomposing a multiclass problem into a series of binary problems



One vs All

build a seperate binary classifier for each class. Choose the class with the highest probability. (r models)

one vs one

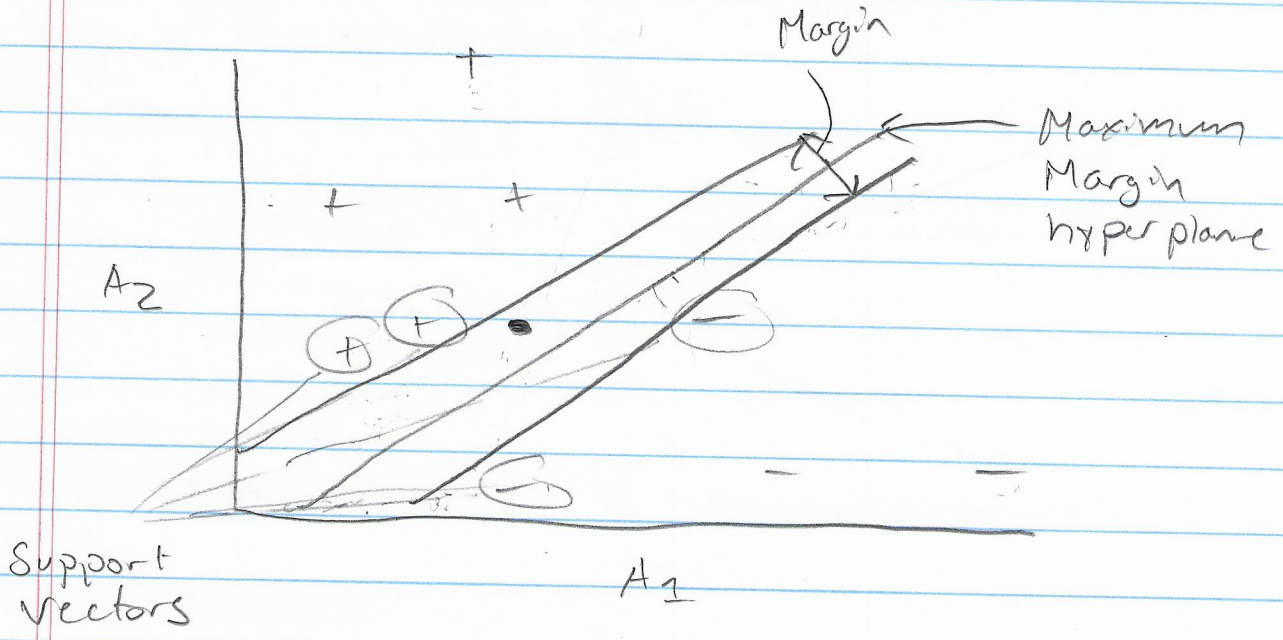
Construct a seperate classifier for each pair of classes. Choose the classifier with the most votes. (r(r-1)/2 models)

Logistic regression, Support Vector Machines, and Neural Nets are all binary classifiers that require either one vs all or one vs one for multiclass problems.

Naive Bayes, Decision Trees, and K-NN are all multiclass models

⑥ Support Vector Machines

Motivation: How do we make the Seperating hyperplane robust?



Basic ideas:

1) only use the data points that are closest to the decision boundry to avoid overfitting (Support vectors)

2) Try to find the seperating hyperplone that maximizes the margin between the two classes (maximum margin hyperplane)