Lecture 9: Classification II: Logistic Regression
Modeling Social Data, Spring 2017
Columbia University

Kaveh Issapour

March 30, 2017

1 Naive Bayes

Here we saw Naive Bayes (independent assumption over all words). We predict class c given features.
Bayes Rule:

$$P(c|x) = p(\bar{x}|c) \times p(c)/p(\bar{x}) \tag{1}$$

The naive part says the following:

$$p(\bar{x}|c) = \prod_j p(x_j|c) \tag{2}$$

and then we have:

$$p(c|\bar{x}) = \prod_j p(x_j|c) \times p(c)/p(\bar{x}) \tag{3}$$

$p(x_j|c)$ models a a coinflip (i.e. Bernoulli)
The word occurrences are coinflips:

$$p(x_j|c) = \theta_{jc}^{x_j}(1-\theta_{jc})^{1-x_j} \tag{4}$$

θ_{jc} predicts the jth word in some class c.

$$log(p(c|x)) = \sum_j log[\theta_{jc}^{x_j}(1-\theta_{jc})^{1-x_j}] + log[\frac{\theta_c}{p(\bar{x})}] \tag{5}$$

$$= \sum_j x_j log\frac{\theta_{jc}}{1-\theta_{jc}} + \sum_j log(1-\theta_{jc}) + +log[\frac{\theta_c}{p(\bar{x})}] \tag{6}$$

The leftmost term is the number of words in the document; the middle term is size of the vocab. we are working with.
We have two cases:

$$log\frac{p(c=1|x)}{p(c=0|\bar{x})} = \sum_j log[\frac{\theta_{j1}(1-\theta_{j0})}{\theta_{j0}(1-\theta_{j1})}] + \sum_j log[\frac{(1-\theta_{j1})}{(1-\theta_{j0})}] + log[\frac{\theta_1}{\theta_0}] \tag{7}$$

Let's look at the difference of the log prob. of both of our cases: Lets define:

$$w_j = log[\frac{\theta_{j1}(1-\theta_{j0})}{\theta_{j0}(1-\theta_{j1})}] \tag{8}$$

So we end up with:

$$log\frac{p(c=1|\bar{x})}{p(c=1|\bar{x})} = \bar{x} \cdot \bar{w_j} + \bar{w_0} \tag{9}$$

Calculate θ_j To do this we will take the derivative of the log-likelihood of the probability of seeing n heads:

$$0 = \frac{n}{\theta} + \frac{N-n}{1-\theta} \tag{10}$$

$$\theta_{j1} = \frac{n_{j1}}{n_1} = \frac{\text{num of spam docs w/ word j}}{\text{num of spam docs}} \tag{11}$$

$$\theta_1 = \frac{n_1}{N} = \frac{\text{num of spam docs}}{\text{num of total docs}} \tag{12}$$

2 Logistic Regression

We shall begin here with the predictor we got above:

$$\log \frac{P}{1-p} = w \cdot x \tag{13}$$

$$p = \frac{1}{1 + e^{-w \cdot x}} \tag{14}$$

We have a set of documents x_i and a set of labels y_i, we know find the model.

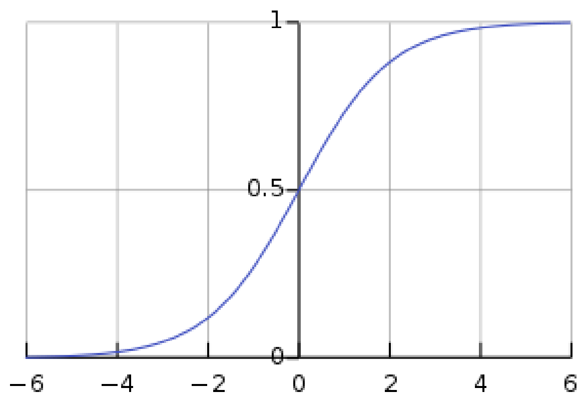

Figure 1: A graph of th Sigmoid Function (from Wikipedia).

$$\text{LL} = P(D|w) = \prod_i p_i^{y_i}(1-p_i)^{1-y_i} \tag{15}$$

Using the above equations and taking the log likelihood we get:

$$L = -\sum_i [y_i w \cdot x - (\log 1 + e^{w \cdot x})] \tag{16}$$

We set the derivative to zero to find the max likelihood and end up with:

$$0 = -\sum_i [y_i \cdot x - (\frac{1}{1 + e^{w \cdot x}})] x_{ik} \tag{17}$$

We use Gradient Descent, since no solution exists:

$$w = w - n \frac{\partial L}{\partial w} \tag{18}$$

It is also:

$$w_k = w_k + n \sum_i [(y_i - p_i) x_{ik}] \tag{19}$$

We can regularize the model as well as follows:

$$L = \sum_i [y_i log_i + (1 - y_i) log(1 - p_i)] + 0.5\lambda ||w||^2 \tag{20}$$

$$\frac{\partial L}{\partial w_k} = \sum_i [(y_i - p_i) x_{ik}] + \lambda w_k \tag{21}$$

We thus end of with:

$$w_k = (1 - n\lambda) w_k + n \sum_i [(y_i - p_i) x_{ik}] \tag{22}$$

3 Evaluation

When evaluating various classifiers we can consider a number of things:

- Accuracy: The fraction of times we predict the correct label.
- Calibration: This is how often an event with predicted probability p occurs.
- Confusion Matrix:

$$p = \frac{1}{1 + e^{-w \cdot x}} \tag{23}$$

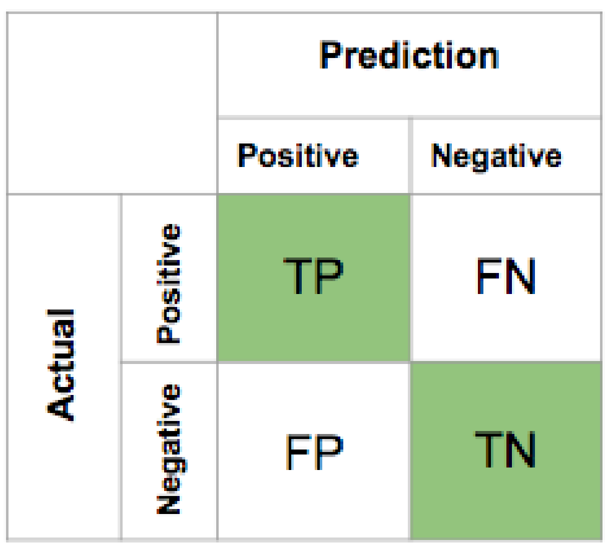

Figure 2: A Confusion Matrix (from WS02).

- Receiver Operating Characteristic (ROC) curve: ROC curve plots the true positive rate and the false positive rate

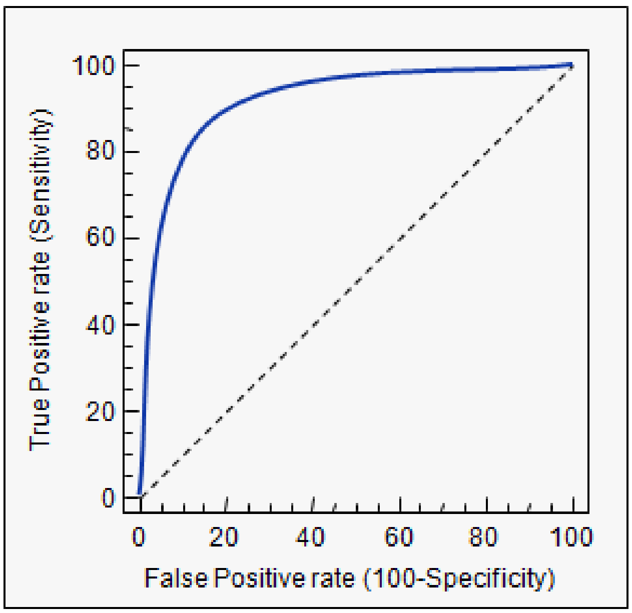

Figure 3: ROC Curve (from Wikipedia).