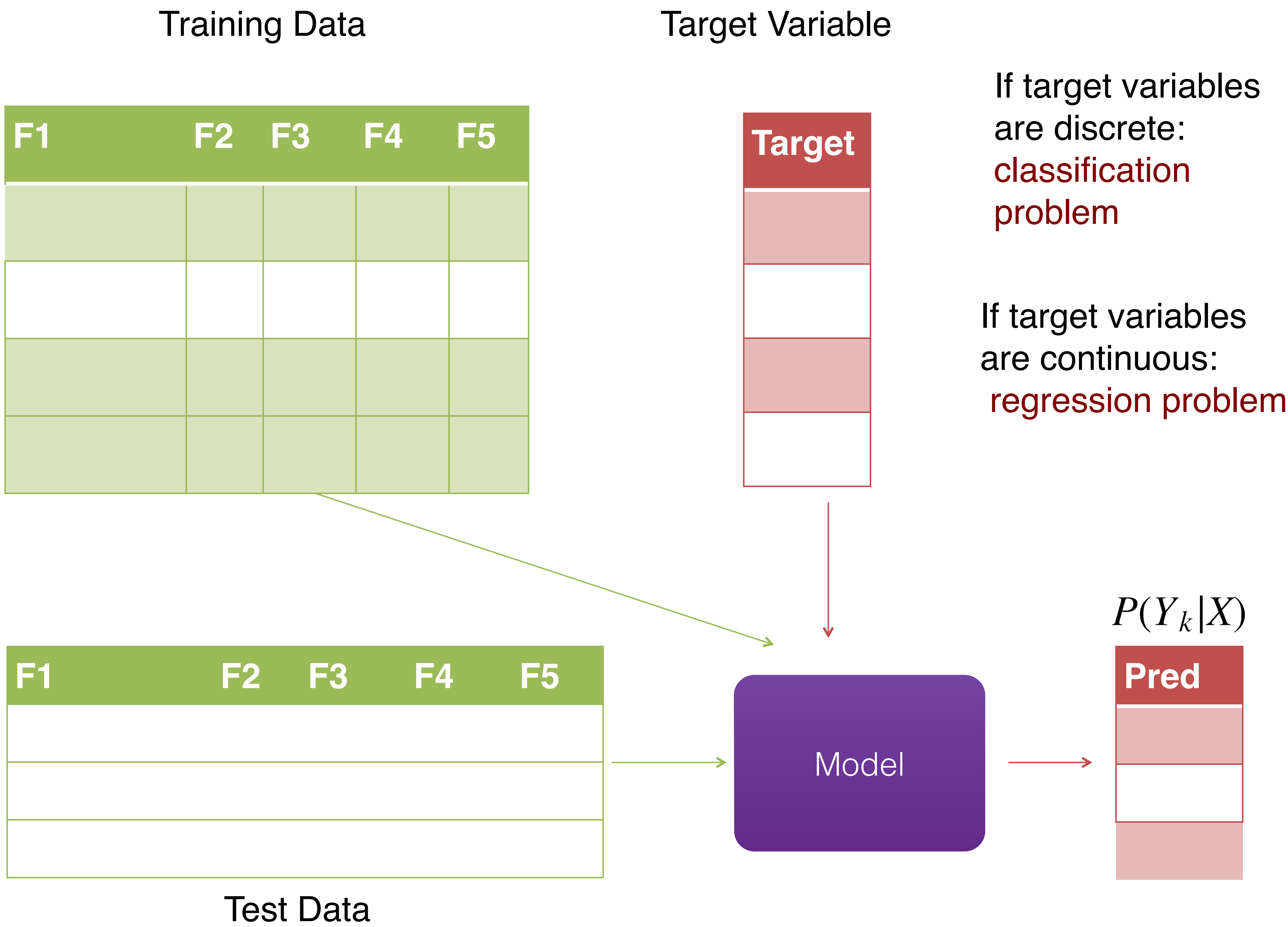
Logistic Regression Review
10-601 Fall 2012
Recitation

September 25, 2012

TA: Selen Uguroglu

Outline

- Decision Theory
- Logistic regression
  - Goal
  - Loss function
  - Inference
  - Gradient Descent


Training Data
Target Variable
F1 F2 F3 F4 F5
Target
If target variables are discrete: classification problem
If target variables are continuous: regression problem
$P(Y_k|X)$
F1 F2 F3 F4 F5
Model
Pred
Test Data

Approach 1: First solve the inference problem of $P(X|Y_k)$ and $P(Y_k)$ separately for each class Y_k. Then use Bayes' theorem to solve:

$$P(Y_k|X) = \frac{P(X|Y_k)P(Y_k)}{P(X)}$$

Approach 2: Infer $P(Y_k|X)$ directly from data

- Generative Models
  - Computationally demanding: requires computing joint distribution over both P(X|Y) and P(Y)
  - Requires large training set for high accuracy
  - Useful for detecting data points that can’t be explained by the current model: anomaly detection/novelty detection
- Discriminative Models
  - Useful if all we want to do is classification

How to perform classification with a discriminative model

We are given the training data, $X = \{<X^1,Y^1>, <X^2,Y^2>, \ldots <X^L,Y^L>\}$ of L examples.

1. Pick a model
2. Estimate the parameters
3. Perform prediction

How to perform classification with a discriminative model

We are given the training data, $X = \{<X^1, Y^1>, <X^2, Y^2>, \ldots <X^L, Y^L>\}$ of L examples.

1. Pick a model
2. Estimate the parameters
3. Perform prediction

Binary logistic regression model

$$P(Y=1|X) = \frac{1}{1+\exp(w_0 + \sum_{i=1}^{n} w_i X_i)}$$

$$P(Y=0|X) = \frac{\exp(w_0 + \sum_{i=1}^{n} w_i X_i)}{1+exp(w_0 + \sum_{i=1}^{n} w_i X_i}$$

Assuming Y can take Boolean values

Multi class logistic regression model

$$P(Y = y_k|X) = \frac{\exp(w_{k0} + \sum_{i=1}^{n} w_{ki}X_i)}{1 + \sum_{j=1}^{K-1} exp(w_{j0} + \sum_{i=1}^{n} w_{ji}X_i)}$$

$$P(Y = y_K|X) = \frac{1}{1 + \sum_{j=1}^{K-1} exp(w_{j0} + \sum_{i=1}^{n} w_{ji}X_i)}$$

One-versus-all classification

How many sets of W's are we predicting?

Side note

To shorten representation, we can add a column of 1's as the 0th feature of X so

$$w_0 + \sum_{i=1}^{n} w_i X_i \qquad \text{becomes} \qquad w^T X$$

Then P(Y=1|X) becomes

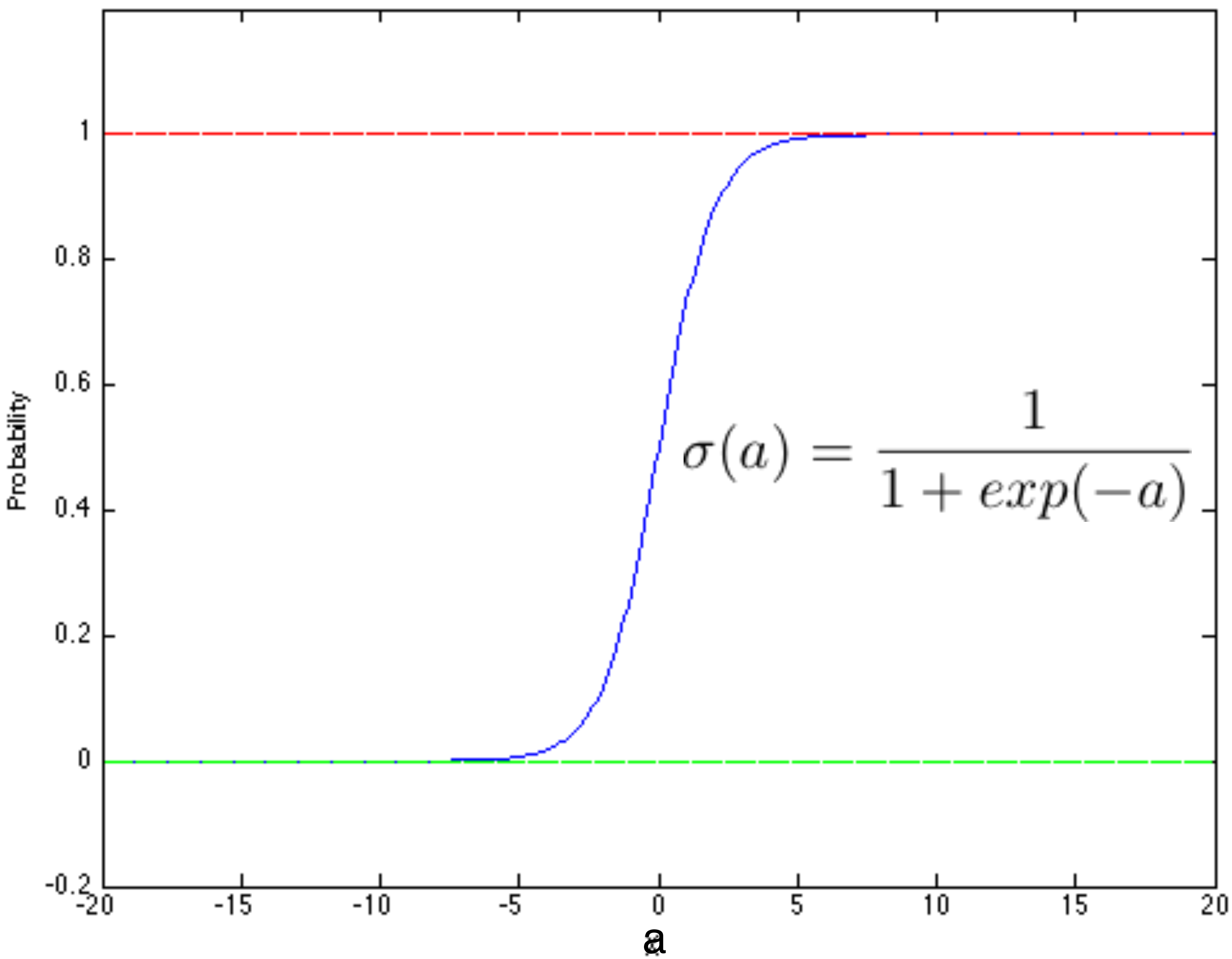
$$P(Y = 1|X) = \frac{1}{1 + exp(-w^T X)}$$

$$P(Y=1|X) = \frac{1}{1+exp(-w^T X)}$$

Sigmoid function $$\sigma(a) = \frac{1}{1+exp(-a)}$$

$$\sigma(w^T X) = \frac{1}{1+exp(-w^T X)}$$

$$\sigma(-a) = 1 - \sigma(a)$$



Monotonically decreases or increases

$$a = \ln\left(\frac{\sigma}{1-\sigma}\right)$$ Logit function

- Range of Logit?
- Relationship with x?

$$\ln \frac{P(Y=0|X)}{P(Y=1|X)} = w_0 + \sum_{i=1}^{n} w_i X_i$$

- Range of odds?

$$\frac{P(Y=0|X)}{P(Y=1|X)} = \exp(w_0 + \sum_{i=1}^{n} w_i X_i)$$

p/(1-p) odds of an event y given x

How to estimate parameters W = <w0,…,wn>?

- Under GNB assumptions
- General case

How to estimate parameters W = <w0,…,wn>?

- Under GNB assumptions
- General case

Let's consider X is a vector of real-valued features

X= <X_1 , X_2, … , X_n>

Xi are conditionally independent given Y

P(Y) ~ Bernoulli(π)

$$P(X_i|Y = y_k) \sim N(\mu_{ik}, \sigma_i)$$

$$P(Y=1|X) = \frac{P(Y=1)P(X|Y=1)}{P(Y=1)P(X|Y=1) + P(Y=0)P(X|Y=0)}$$

$$P(Y=1|X) = \frac{1}{1 + \frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}}$$

$$P(Y=1|X) = \frac{1}{1 + \exp\left(\ln \frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}\right)}$$

Using conditional independence assumption and priors

$$P(Y=1|X) = \frac{1}{1 + \exp\left(\ln \frac{1-\pi}{\pi} + \sum_i \ln \frac{P(X_i|Y=0)}{P(X_i|Y=1)}\right)}$$

$$P(Y=1|X) = \frac{1}{1+\exp\left(\ln\frac{1-\pi}{\pi} + \sum_i \ln\frac{P(X_i|Y=0)}{P(X_i|Y=1)}\right)}$$

Since variables have Gaussian distribution:

$$\sum_i \left(\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}X_i + \frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}\right)$$

$$P(Y=1|X) = \frac{1}{1+exp(w_0+\sum_{i=1}^n w_i X_i)}$$

How to estimate parameters W = <w0,…,wn>?

- Under GNB assumptions:

1. Variables Xi are conditionally independent given Y
2. $P(X_i|Y = y_k) \sim N(\mu_{ik}, \sigma_i)$

- General case

A. MLE

B. MAP

Estimating parameters with MLE

$$W \leftarrow \arg\max_W \prod_l P(Y^l | X^l, W)$$

Conditional likelihood

What's data likelihood?

Let's write the log conditional likelihood first

$$L(W) = \prod_l P(Y^l|X^l, W)$$

Is there a conditional independence assumption here?

$$l(W) = \ln \prod_l P(Y^l|X^l, W)$$

Taking the log

$$l(W) = \sum_l \ln P(Y^l|X^l, W)$$

$$l(W) = \sum_l Y^l \ln P(Y^l = 1|X^l, W) + (1 - Y^l) \ln P(Y^l = 0|X^l, W)$$

$$l(W) = \sum_l Y^l \ln(w_0 + \sum_i^n w_i X_i^l) + \ln(1 + \exp(w_0 + \sum_i^n w_i X_i^l))$$

Objective function that I want to maximize

$$l(W) = \sum_l Y^l \ln(w_0 + \sum_i^n w_i X_i^l) + \ln(1 + \exp(w_0 + \sum_i^n w_i X_i^l))$$

One problem: No closed form solution!!!

$$l(W) = \sum_l Y^l \ln(w_0 + \sum_i^n w_i X_i^l) + \ln(1 + \exp(w_0 + \sum_i^n w_i X_i^l))$$

Take partial derivatives with respect to wi

$$\frac{\partial l(W)}{\partial w_i} = \sum_l X_i^l (Y^l - \hat{P}(Y^l = 1 | X^l, W))$$

$$w_i \leftarrow w_i + \eta \sum_l X_i^l (Y^l - \hat{P}(Y^l = 1 | X^l, W))$$

Step size

Gradient descent

First order optimization

Taking steps to the direction of the negative gradient of the function

Suppose we want to minimize F(x) which is defined and differentiable at point z

F(x) decreases the fastest I start from point z and go to the direction of the negative gradient of F(z)

$$z_{(n+1)} = z_n - \eta \nabla F(z_n)$$

Gradient ascent

First order optimization

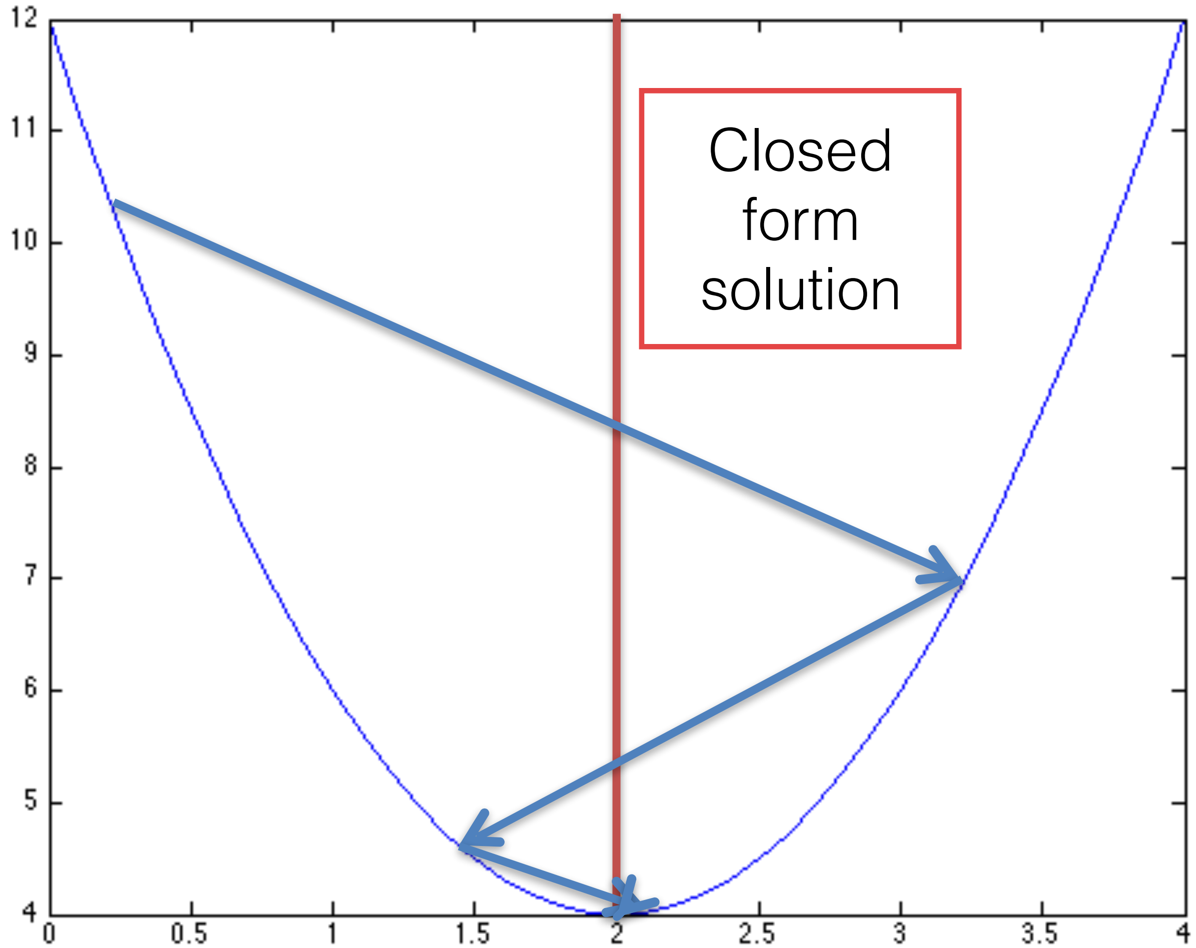
Taking steps to the direction of the positive gradient of the function

Suppose we want to maximize F(x) which is defined and differentiable at point z

F(x) increases the fastest I start from point z and go to the direction of the positive gradient of F(z)

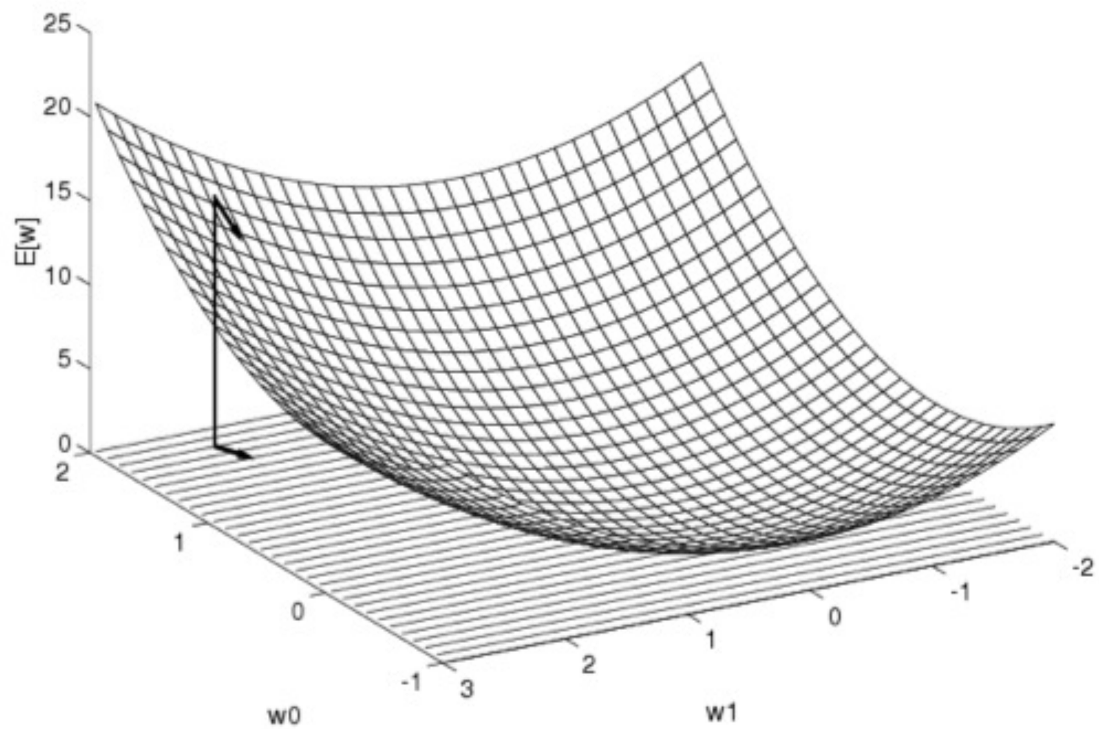
$$z_{(n+1)} = z_n + \eta \nabla F(z_n)$$

Gradient descent



- The function we want to minimize is the parabola show in light blue
- The closed form solution is available
- Starting from a random point on parabola gradient descent takes steps to reach a local minima

Gradient Descent



Gradient

$$\nabla E[\vec{w}] \equiv \left[\frac{\partial E}{\partial w_0}, \frac{\partial E}{\partial w_1}, \cdots \frac{\partial E}{\partial w_n}\right]$$

Training rule:

$$\Delta \vec{w} = -\eta \nabla E[\vec{w}]$$

i.e.,

$$\Delta w_i = -\eta \frac{\partial E}{\partial w_i}$$

$$w_i \leftarrow w_i + \eta \sum_l X_i^l (Y^l - \hat{P}(Y^l = 1 | X^l, W))$$

- Sum over all training examples
- What if I have a large training data
- Summation after each iteration
- Slow!!!!

$$F(z) = \sum_{i=1}^{n} F_i(z)$$

Function I want to minimize

$$z \leftarrow z - \eta \sum_{i=1}^{n} \nabla F_i(z)$$

Batch gradient descent

$$z \leftarrow z - \eta \nabla F_i(z)$$

Stochastic gradient descent

Stochastic gradient descent update rule?

$$w_i \leftarrow w_i + \eta(X_i^l(Y^l - \hat{(}P(Y^l = 1|X^l, W))$$

How to pick the training instance?

- How to pick the learning rate?
- Suppose the features have varying ranges. Would that be a problem?

Stochastic gradient descent

- Faster convergence when the training data is large

- Learning rate should be low otherwise there is a risk of going back and forth

- High accuracy is hard to reach

Batch gradient descent

- Slow convergence when the training data is large
- Guaranteed to reach to a local minimum under certain conditions

How to estimate parameters W = <w0,…,wn>?

- Under GNB assumptions
- General case

  A. MLE

  B. MAP

Estimating parameters with MAP

$$W \leftarrow \arg\max_W \ln P(W) \prod_l P(Y^l|X^l, W)$$

Assume P(W) has a Gaussian distribution with zero mean identity covariance

$$w_i \leftarrow w_i - \eta\lambda w_i + \eta \sum_l X_i^l (Y^l - \hat{P}(Y^l = 1|X^l, W))$$

From the prior

$$w_i \leftarrow w_i - \eta\lambda w_i + \eta \sum_l X_i^l (Y^l - \hat{P}(Y^l = 1 | X^l, W$$

- Defining parameters on W corresponds to regularization
- Pushes parameters towards 0
- Avoids large weights and over fitting

Generative versus Discriminative

Generative

- Assumes a functional form of P(X|Y) and P(Y)
- Estimates P(X|Y) and P(Y) from data, uses them to calculate P(Y|X)

Discriminative

- Assumes a functional form of P(Y|X)
- Estimates P(Y|X) directly from the data

$$P(X_1, X_2, \ldots, X_n|Y) = \prod_i^n P(X_i|Y)$$

Performance as training data reaches infinity

$$P(X_i|Y = y_k) \sim N(\mu_{ik}, \sigma_i)$$

Decision Surfaces

Naive Bayes

Gaussian Naive Bayes

Logistic Regression

Things to think about

- Overfitting in LR
- Does LR make any assumptions on P(X|Y)?
- GNB with class independent variances is generative equivalent of LR under GNB assumptions
- What's the objective function of LR? Can we reach global optimum?

Questions?

Acknowledgements

- Some slides are taken from Tom Mitchell's 10-601 lecture notes