

# Logistic Regression

Machine Learning – 10701/15781  
 Carlos Guestrin  
 Carnegie Mellon University  
 September 24<sup>th</sup>, 2007

## Generative v. Discriminative classifiers – Intuition

- **Want to Learn:**  $h: X \mapsto Y$   $Y \in \{1, 2, 3, \dots, k\}$ 
  - $X$  – features
  - $Y$  – target classes
- **Bayes optimal classifier** –  $P(Y|X)$
- **Generative classifier**, e.g., Naive Bayes:
  - Assume some **functional form for  $P(X|Y), P(Y)$** 
    - Estimate parameters of  $P(X|Y), P(Y)$  directly from training data
    - Use Bayes rule to calculate  $P(Y|X=x) = \frac{P(Y, X=x)}{P(X=x)}$
    - This is a '**generative**' model
      - Indirect computation of  $P(Y|X)$  through Bayes rule
      - But, **can generate a sample of the data**,  $P(X) = \sum_y P(y) P(X|y)$
- **Discriminative classifiers**, e.g., Logistic Regression:
  - Assume some **functional form for  $P(Y|X)$**
  - Estimate parameters of  $P(Y|X)$  directly from training data
  - This is the '**discriminative**' model
    - Directly learn  $P(Y|X)$
    - But **cannot obtain a sample of the data**, because  $P(X)$  is not available

generate spem: sample (or set)  $P(Y = spam)$   
 sample words:  $P(X|Y=spam)$   
 eg. NB:  $P(X|Y) = \prod P(X_i|Y)$   $P(Y)$   
 exactly  
 at classification time: input  $x$  answer  $P(Y|X=x)$

learn  $P(Y, X)$

# Logistic Regression

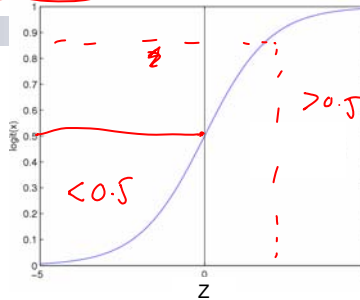
Logistic function (or Sigmoid):

$$\frac{1}{1 + \exp(-z)}$$

Learn  $P(Y|X)$  directly!

- Assume a particular functional form
- Sigmoid applied to a linear function of the data:

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



Features can be discrete or continuous!

3

# Understanding the sigmoid

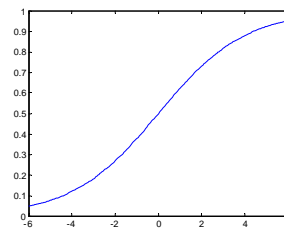
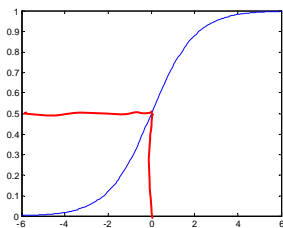
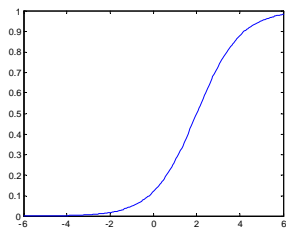
$$g(w_0 + \sum_i w_i x_i) = \frac{1}{1 + e^{w_0 + \sum_i w_i x_i}}$$

constant  $w_0 + w_1 x_1$

$w_0 = -2, w_1 = -1$

$w_0 = 0, w_1 = -1$

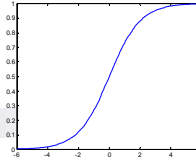
$w_0 = 0, w_1 = -0.5$



©Carlos Guestrin 2005-2007

4

# Logistic Regression – a Linear classifier



$g(w_0 + \sum_{i=1}^n w_i x_i) = \frac{1}{1 + e^{w_0 + \sum_{i=1}^n w_i x_i}}$

$w_0 + \sum_{i=1}^n w_i x_i < 0$   
 $x_i: g > 0.5$  → true

$w_0 + \sum_{i=1}^n w_i x_i = 0$   
 $g = 0.5 = \frac{1}{1+1}$

$w_0 + \sum_{i=1}^n w_i x_i > 0$   
 $x_i: g < 0.5$  → false

*n-dimensional space*

©Carlos Guestrin 2005-2007

5

## Very convenient!

$\ln 1 = 0$

$P(Y = 1 | X = \langle X_1, \dots, X_n \rangle) = \frac{1}{1 + \exp(w_0 + \sum_i w_i X_i)}$

implies  $= 1 - P(Y = 0 | X, w)$

$P(Y = 0 | X = \langle X_1, \dots, X_n \rangle) = \frac{\exp(w_0 + \sum_i w_i X_i)}{1 + \exp(w_0 + \sum_i w_i X_i)}$

implies

$\frac{P(Y = 0 | X)}{P(Y = 1 | X)} = \exp(w_0 + \sum_i w_i X_i)$

implies

$\ln \frac{P(Y = 0 | X)}{P(Y = 1 | X)} = w_0 + \sum_i w_i X_i < 0$  return  $Y = 1$

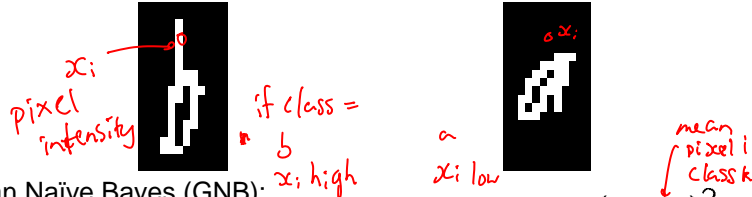
linear classification rule!

©Carlos Guestrin 2005-2007

6

# What if we have continuous $X_i$ ?

Eg., character recognition:  $X_i$  is  $i^{\text{th}}$  pixel



Gaussian Naïve Bayes (GNB):

$$P(X_i = x | Y = y_k) = \frac{1}{\sigma_{ik} \sqrt{2\pi}} e^{-\frac{(x - \mu_{ik})^2}{2\sigma_{ik}^2}}$$

Handwritten notes:  $\mu_{ik}$  (mean pixel i class k),  $\sigma_{ik}^2$  (variance pixel i class k),  $y_k$  (label)

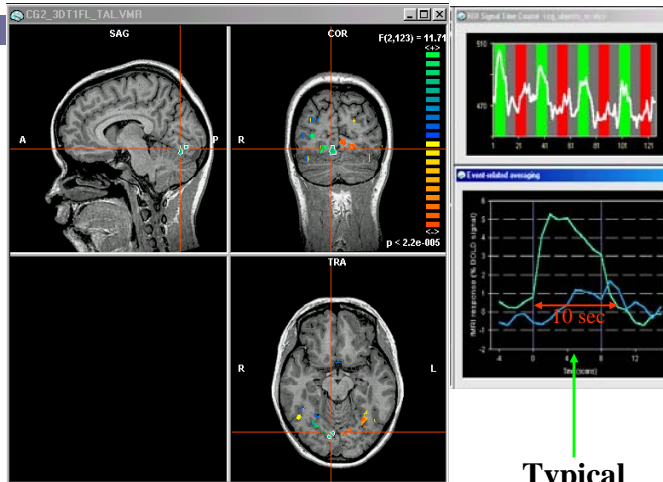
Sometimes assume variance

- is independent of Y (i.e.,  $\sigma_i$ ),
- or independent of  $X_i$  (i.e.,  $\sigma_k$ )
- or both (i.e.,  $\sigma$ )

## Example: GNB for classifying mental states

[Mitchell et al.]

~1 mm resolution  
 ~2 images per sec.  
 15,000 voxels/image  
 non-invasive, safe  
 measures Blood Oxygen Level Dependent (BOLD) response



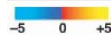
Typical impulse response

## Learned Bayes Models – Means for $P(\text{BrainActivity} | \text{WordCategory})$

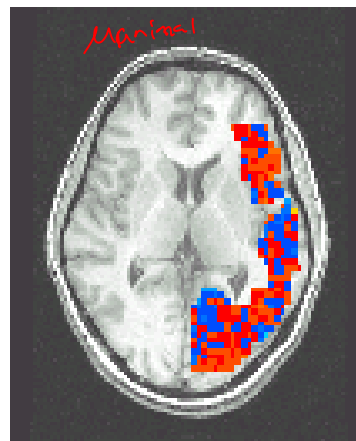
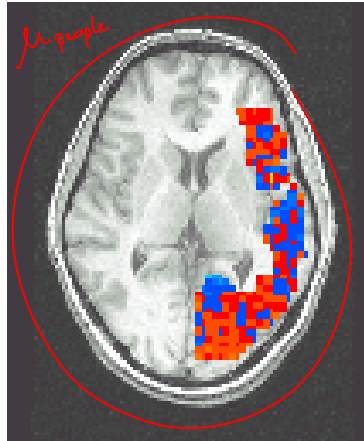
[Mitchell et al.]

Pairwise classification accuracy: 85%

People words



Animal words



©Carlos Guestrin 2005-2007

9

## Logistic regression v. Naïve Bayes

- Consider learning  $f: X \rightarrow Y$ , where
  - $X$  is a vector of real-valued features,  $\langle X_1 \dots X_n \rangle$
  - $Y$  is boolean

- Could use a Gaussian Naïve Bayes classifier

- assume all  $X_i$  are conditionally independent given  $Y$
  - model  $P(X_i | Y = y_k)$  as Gaussian  $N(\mu_{ik}, \sigma_i)$
  - model  $P(Y)$  as Bernoulli( $\theta, 1-\theta$ )
- variance only depends on  $x_i$  on pixel  $i$ , not on class*

- What does that imply about the form of  $P(Y|X)$ ?

$$P(Y = 1 | X = \langle X_1, \dots, X_n \rangle) = \frac{1}{1 + \exp(w_0 + \sum_i w_i X_i)}$$

**Cool!!!!**

©Carlos Guestrin 2005-2007

10

## Derive form for $P(Y|X)$ for continuous $X_i$

$e^{\ln x} = x$

$P(Y=1|X) = \frac{1}{1 + e^{\ln \theta + \sum_i w_i x_i}}$

*Bayes rule*

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

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

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

$$= \frac{1}{1 + \exp(\underbrace{\ln \frac{1-\theta}{\theta}}_{\text{looks like } w_0} + \underbrace{\sum_i \ln \frac{P(X_i|Y=0)}{P(X_i|Y=1)}}_{\text{independent of } x_i})}$$

*looks like  $w_0$   
independent of  $x_i$*

11

©Carlos Guestrin 2005-2007

## Ratio of class-conditional probabilities

$\ln \frac{1}{e} = -1$

$P(Y=1|X) = \frac{1}{1 + e^{\ln \theta + \sum_i w_i x_i}}$

*i indexes over features*

$$\ln \frac{P(X_i|Y=0)}{P(X_i|Y=1)} =$$

$$P(X_i = x_i | Y = y_k) = \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{(x_i - \mu_{ik})^2}{2\sigma_i^2}}$$

*var doesn't depend on class k*

$$\ln \frac{\frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{(x_i - \mu_{i0})^2}{2\sigma_i^2}}}{\frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{(x_i - \mu_{i1})^2}{2\sigma_i^2}}} =$$

$$\frac{-(x_i - \mu_{i0})^2}{2\sigma_i^2} + \frac{(x_i - \mu_{i1})^2}{2\sigma_i^2}$$

$$= \frac{(\mu_{i0} - \mu_{i1}) x_i}{\sigma_i^2} + \frac{(\mu_{i1}^2 - \mu_{i0}^2)}{2\sigma_i^2}$$

$$= \frac{-x_i^2 + 2x_i\mu_{i0} - \mu_{i0}^2 + x_i^2 - 2\mu_{i1}x_i + \mu_{i1}^2}{2\sigma_i^2}$$

12

©Carlos Guestrin 2005-2007

## Derive form for $P(Y|X)$ for continuous $X_i$

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

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

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

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

$$w_0 = \ln \frac{1-\theta}{\theta} + \sum_i \frac{\mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2}$$

©Carlos Guestrin 2005-2007

13

## Gaussian Naïve Bayes v. Logistic Regression

**Set of Gaussian Naïve Bayes parameters**  
(feature variance independent of class label)

**Set of Logistic Regression parameters**

*transform into parameterization of LR*

*transform to NB, but not all w's*

- Representation equivalence
  - **But only in a special case!!!** (GNB with class-independent variances)
- But what's the difference???
- **LR makes no assumptions about  $P(X|Y)$  in learning!!!**
- **Loss function!!!**
  - Optimize different functions → Obtain different solutions

*does not assume independence*  
*assume form for  $P(Y|X)$*

©Carlos Guestrin 2005-2007

14

# Logistic regression for more than 2 classes

- Logistic regression in more general case, where  $Y \in \{Y_1 \dots Y_R\}$ : learn  $R-1$  sets of weights

4 class: 3 sets of parameters

$$P(Y=1|X, w_1) \propto e^{w_{10} + \sum_i w_{1i} X_i}$$

$$P(Y=2|X, w_2) \propto e^{w_{20} + \sum_i w_{2i} X_i}$$

$$\vdots$$

$$P(Y=R-1|X, w_{R-1}) \propto e^{w_{R-1,0} + \sum_i w_{R-1,i} X_i}$$

$$P(Y=R|X) = 1 - \sum_{j=1}^{R-1} P(Y=j|X) \propto 1 - \sum_{j=1}^{R-1} e^{w_{j0} + \sum_i w_{ji} X_i}$$

15

©Carlos Guestrin 2005-2007

# Logistic regression more generally

- Logistic regression in more general case, where  $Y \in \{Y_1 \dots Y_R\}$ : learn  $R-1$  sets of weights

for  $k < R$

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

for  $k=R$  (normalization, so no weights for this class)

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

Features can be discrete or continuous!

$X_i = \text{gender in 1070?}$

A=1  
B=2  
C=3

16

©Carlos Guestrin 2005-2007

# Loss functions: Likelihood v. Conditional Likelihood

- Generative (Naïve Bayes) Loss function:

**Data likelihood**

$$\ln P(\mathcal{D} | \mathbf{w}) = \sum_{j=1}^N \ln P(\mathbf{x}^j, y^j | \mathbf{w})$$

$$= \sum_{j=1}^N \ln P(y^j | \mathbf{x}^j, \mathbf{w}) + \sum_{j=1}^N \ln P(\mathbf{x}^j | \mathbf{w})$$

- Discriminative models cannot compute  $P(\mathbf{x} | \mathbf{w})!$
- But, discriminative (logistic regression) loss function:

**Conditional Data Likelihood**

$$\ln P(\mathcal{D}_Y | \mathcal{D}_X, \mathbf{w}) = \sum_{j=1}^N \ln P(y^j | \mathbf{x}^j, \mathbf{w})$$

- Doesn't waste effort learning  $P(X)$  – focuses on  $P(Y|X)$  all that matters for classification

$\mathcal{D} = \langle \mathbf{x}^j, y^j \rangle_{j=1 \dots N}$

$P(\mathbf{x}, y | \mathbf{w}) = P(y | \mathbf{x}, \mathbf{w}) \cdot P(\mathbf{x} | \mathbf{w})$

for generating data not important for classification

discriminative likelihood

$i = \text{training example}$   
 $y^j = 1$  if spam  
 $= 0$  if not spam  
 $\mathbf{x}^j = \text{list of words}$   
 17 in 1-11

©Carlos Guestrin 2005-2007

## Expressing Conditional Log Likelihood

$\max_{\mathbf{w}} l(\mathbf{w}) \equiv \sum_j \ln P(y^j | \mathbf{x}^j, \mathbf{w})$

$$P(Y = 0 | \mathbf{X}, \mathbf{w}) = \frac{1}{1 + \exp(w_0 + \sum_i w_i X_i)}$$

$$P(Y = 1 | \mathbf{X}, \mathbf{w}) = \frac{\exp(w_0 + \sum_i w_i X_i)}{1 + \exp(w_0 + \sum_i w_i X_i)}$$

$j$ th component:  $P(Y = \text{spam} | \mathbf{x}^j, \mathbf{w})$

if  $j$ th was spam  
 $P(Y = \text{not spam} | \mathbf{x}^j, \mathbf{w})$   
 if  $j$ th was not spam

$$l(\mathbf{w}) = \sum_j [y^j \ln P(y = 1 | \mathbf{x}^j, \mathbf{w}) + (1 - y^j) \ln P(y = 0 | \mathbf{x}^j, \mathbf{w})]$$

if  $y^j = 1$  :  $\ln P(y = 1 | \mathbf{x}^j, \mathbf{w}) + 0$

if  $y^j = 0$  :  $0 + \ln P(y = 0 | \mathbf{x}^j, \mathbf{w})$

$$l(\mathbf{w}) = \sum_j \left[ y^j [w_0 + \sum_i w_i x_i - \ln(1 + e^{w_0 + \sum_i w_i x_i})] + (1 - y^j) [-\ln(1 + e^{w_0 + \sum_i w_i x_i})] \right]$$

©Carlos Guestrin 2005-2007

18

# Maximizing Conditional Log Likelihood

$$\begin{aligned}
 P(Y = 0|X, W) &= \frac{1}{1 + \exp(w_0 + \sum_i w_i X_i)} \\
 P(Y = 1|X, W) &= \frac{\exp(w_0 + \sum_i w_i X_i)}{1 + \exp(w_0 + \sum_i w_i X_i)} \\
 \max_{\mathbf{w}} l(\mathbf{w}) &\equiv \ln \prod_j P(y^j | \mathbf{x}^j, \mathbf{w}) \\
 &= \sum_j \left[ \underbrace{y^j (w_0 + \sum_i w_i x_i^j)}_{\text{linear part}} - \ln(1 + \exp(w_0 + \sum_i w_i x_i^j)) \right]
 \end{aligned}$$

**Good news:**  $l(\mathbf{w})$  is concave function of  $\mathbf{w}$  → no locally optimal solutions

**Bad news:** no closed-form solution to maximize  $l(\mathbf{w})$

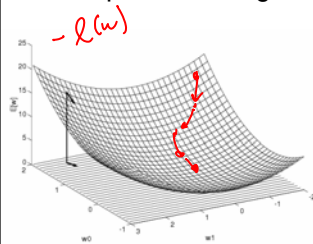
**Good news:** concave functions easy to optimize



# Optimizing concave function – Gradient ascent

*(Conjugate G.D.)  
better.*

- Conditional likelihood for Logistic Regression is concave → Find optimum with gradient ascent



**Gradient:**  $\nabla_{\mathbf{w}} l(\mathbf{w}) = \left[ \frac{\partial l(\mathbf{w})}{\partial w_0}, \dots, \frac{\partial l(\mathbf{w})}{\partial w_n} \right]^T$

**Update rule:**  $\Delta \mathbf{w} = \eta \nabla_{\mathbf{w}} l(\mathbf{w})$

*step size*  
**Learning rate,  $\eta > 0$**

*0.01*

$$w_i^{(t+1)} \leftarrow w_i^{(t)} + \eta \frac{\partial l(\mathbf{w})}{\partial w_i}$$

- Gradient ascent is simplest of optimization approaches
  - e.g., Conjugate gradient ascent much better (see reading)

# Maximize Conditional Log Likelihood:

## Gradient ascent

$$P(Y=0|X,W) = \frac{1}{1 + \exp(w_0 + \sum_i w_i X_i)}$$

$$P(Y=1|X,W) = \frac{\exp(w_0 + \sum_i w_i X_i)}{1 + \exp(w_0 + \sum_i w_i X_i)}$$

$$l(w) = \sum_j y^j (w_0 + \sum_i w_i x_i^j) - \ln(1 + \exp(w_0 + \sum_i w_i x_i^j))$$

$$\frac{\partial l(w)}{\partial w_i} = \sum_j [y^j x_i^j - \frac{\partial}{\partial w_i} \ln(1 + e^{w_0 + \sum_i w_i x_i^j})]$$

$$= \sum_j [y^j x_i^j - \frac{x_i^j e^{w_0 + \sum_i w_i x_i^j}}{1 + e^{w_0 + \sum_i w_i x_i^j}}]$$

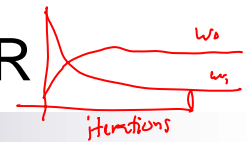
$p(Y=1|X,w)$

$= \sum_j x_i^j [y^j - p(Y=1|X,w)]$

if  $j$ th example is positive: if  $x_i^j$  is positive want to make  $w_i$  large  
 if  $j$ th " is negative: if  $x_i^j$  is positive want to make  $w_i$  small

©Carlos Guestrin 2005-2007

# Gradient Descent for LR



Gradient ascent algorithm: iterate until change  $< \epsilon$

$$w_0^{(t+1)} \leftarrow w_0^{(t)} + \eta \sum_j [y^j - \tilde{P}(Y^j = 1 | x^j, w^{(t)})]$$

*Handwritten notes:*  $w^{(t)}$ :  $w$  at  $t$ th iteration;  $x_0^j = 1$ ;  $x_0^j$  no  $x_0^j$

For  $i = 1 \dots n$ ,

$$w_i^{(t+1)} \leftarrow w_i^{(t)} + \eta \sum_j x_i^j [y^j - \tilde{P}(Y^j = 1 | x^j, w^{(t)})]$$

repeat

$$\frac{e^{w_0 + \sum_i w_i x_i^j}}{1 + e^{w_0 + \sum_i w_i x_i^j}}$$

©Carlos Guestrin 2005-2007