# CSE 446
# Logistic Regression
Winter 2012

Dan Weld

Some slides from Carlos Guestrin, Luke Zettlemoyer

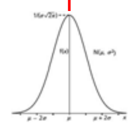
## Gaussian Naïve Bayes

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

$$P(Y \mid \mathbf{X}) \propto P(\mathbf{X} \mid Y)\, P(Y)$$

$$P(X_i = x \mid Y = y_k) = N(\mu_{ik}, \sigma_{ik})$$

$$N(\mu_{ik}, \sigma_{ik}) = \frac{1}{\sigma_{ik}\sqrt{2\pi}}\, e^{\frac{-(x-\mu_{ik})^2}{2\sigma_{ik}^2}}$$



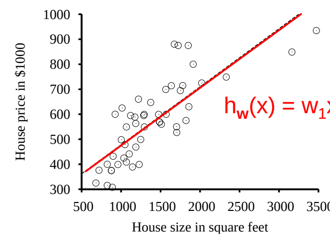
## **Generative** *vs.* **Discriminative** Classifiers

- **Want to Learn**: h:**X** ↦ Y
  - **X** – features
  - Y – target classes
- **Bayes optimal classifier** – P(Y|**X**)
- **Generative classifier**, e.g., Naïve 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)
  - This is a ***'generative'* model**
    - **Indirect** computation of P(Y|X) through Bayes rule
    - As a result, **can also 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

$$P(Y \mid \mathbf{X}) \propto P(\mathbf{X} \mid Y)\, P(Y)$$



## Univariate Linear Regression

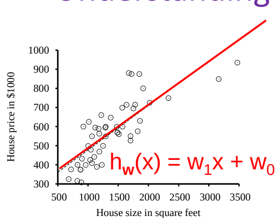



$h_w(x) = w_1 x + w_0$

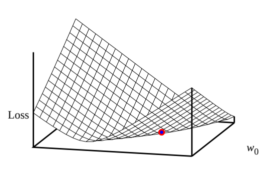
$$Loss(h_w) = \sum_{j=1}^{n} L_2(y_j, h_w(x_j)) = \sum_{j=1}^{n} (y_j - h_w(x_j))^2 = \sum_{j=1}^{n} (y_j - (w_1 x_j + w_0))^2$$



## Understanding Weight Space



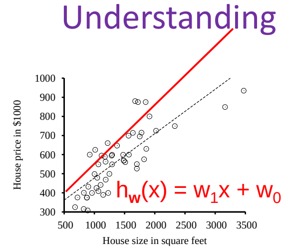

$h_w(x) = w_1 x + w_0$



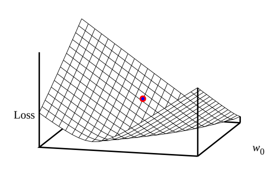

$$Loss(h_w) = \sum_{j=1}^{n} (y_j - (w_1 x_j + w_0))^2$$



## Understanding Weight Space



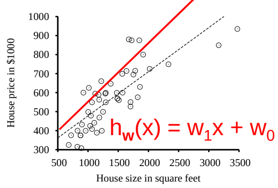

$h_w(x) = w_1 x + w_0$

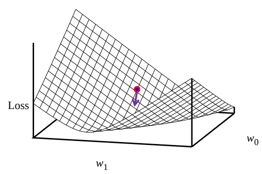



$$Loss(h_w) = \sum_{j=1}^{n} (y_j - (w_1 x_j + w_0))^2$$

## Finding Minimum Loss

Argmin$_w$ Loss($h_w$)



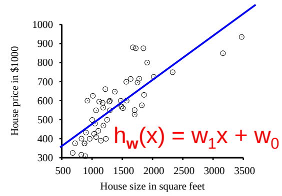

$$\text{Loss}(h_w) = \sum_{j=1}^{n}(y_j - (w_1x_j+w_0))^2$$

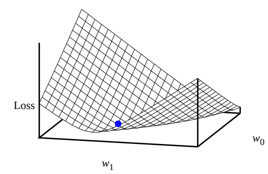
$$\frac{\partial}{\partial w_0}\text{Loss}(h_w) = 0 \qquad \frac{\partial}{\partial w_1}\text{Loss}(h_w) = 0$$



## Unique Solution!

Argmin$_w$ Loss($h_w$)



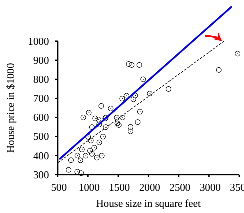



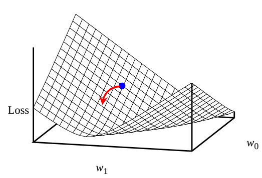

$$w_1 = \frac{N\sum(x_jy_j)-(\sum x_j)(\sum y_j)}{N\sum(x_j^2)-(\sum x_j)^2}$$

$$w_0 = (\sum(y_j) - w_1(\sum x_j)/N$$

## Could also Solve Iteratively

Argmin$_w$ Loss($h_w$)




**w** = any point in weight space
Loop until convergence
For each $w_i$ in w do
$w_i := w_i - \alpha\frac{\partial}{\partial w_i}\text{Loss}(\mathbf{w})$

## Multivariate Linear Regression

$h_w(\mathbf{x_j}) = w_0 + \sum w_i\, x_{j,i} \quad = \sum w_i\, x_{j,i} \quad = w^Tx_j$

Argmin$_w$ Loss($h_w$)

Unique Solution = $(x^Tx)^{-1}x^Ty$

Problem….



## Overfitting

Regularize!!

Penalize high weights

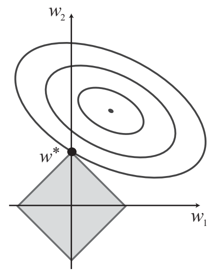
$$\text{Loss}(h_w) = \sum_{j=1}^{n}(y_j - (w_1x_j+w_0))^2 \quad +\lambda\sum_{i=1}^{k} w_i^2$$

Alternatively….

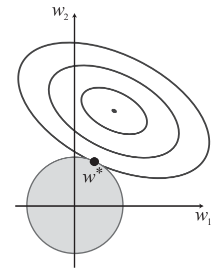
$$\text{Loss}(h_w) = \sum_{j=1}^{n}(y_j - (w_1x_j+w_0))^2 \quad +\lambda\sum_{i=1}^{k} |w_i|$$



## Regularization

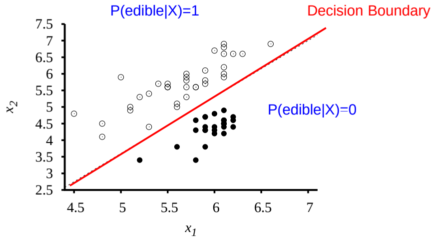



L1




L2

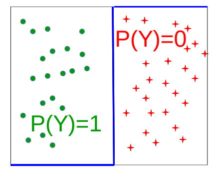
## Back to Classification




13

## Logistic Regression

- Learn P(Y|**X**) directly!
  - Assume a particular functional form
  - ☹ ***Not differentiable…***

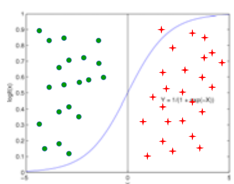



14

## Logistic Regression

- Learn P(Y|**X**) directly!
  - Assume a particular functional form
  - Logistic Function
    - Aka Sigmoid

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

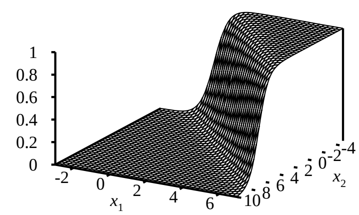


15

## Logistic Function in n Dimensions

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

Sigmoid applied to a linear function of the data:



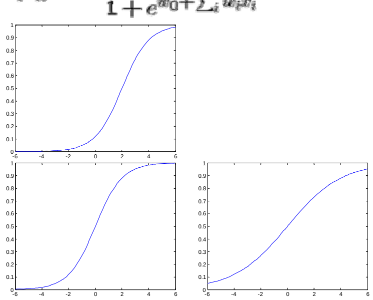

**Features can be discrete or continuous!**

16

## Understanding Sigmoids

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

$w_0$=-2, $w_1$=-1



$w_0$=0, $w_1$=-1

$w_0$=0, $w_1$= -0.5

## Very convenient!

$$P(Y=1|X=<X_1,...X_n>)=\frac{1}{1+exp(w_0+\sum_i w_i X_i)}$$

implies

$$P(Y=0|X=<X_1,...X_n>)=\frac{exp(w_0+\sum_i w_i X_i)}{1+exp(w_0+\sum_i w_i X_i)}$$

implies

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

linear classification rule!

implies

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

18

## Logistic regression more generally

Logistic regression in more general case, where $Y \in \{y_1,\ldots,y_R\}$

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!**



## Loss Functions: Likelihood vs. Conditional Likelihood

Generative (Naïve Bayes) Loss function: **Data likelihood**

$$\begin{aligned} \ln P(\mathcal{D} \mid \mathbf{w}) &= \sum_{j=1}^{N} \ln P(\mathbf{x}^j, y^j \mid \mathbf{w}) \\ &= \sum_{j=1}^{N} \ln P(y^j \mid \mathbf{x}^j, \mathbf{w}) + \sum_{j=1}^{N} \ln P(\mathbf{x}^j \mid \mathbf{w}) \end{aligned}$$

Discriminative (Logistic Regr.) Loss funct: **Conditional Data Likelihood**

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

Discriminative models *can't* compute P($\mathbf{x}^j$|**w**)!

Or, … "They don't *waste effort* learning P(**X**)"

Focus only on P(Y|**X**) - all that matters for classification



## Expressing Conditional Log Likelihood

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

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

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

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

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

$$l(\mathbf{w}) = \sum_j \underset{\text{1 when correct answer is 1}}{y^j} \underbrace{\ln P(y^j = 1|\mathbf{x}^j, \mathbf{w})}_{\text{Probability of predicting 1}} + \underset{\text{1 when correct answer is 0}}{(1 - y^j)} \underbrace{\ln P(y^j = 0|\mathbf{x}^j, \mathbf{w})}_{\text{Probability of predicting 0}}$$



## Expressing Conditional Log Likelihood

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

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

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

$$\begin{aligned} l(\mathbf{w}) &= \sum_j y^j \ln P(y^j = 1|\mathbf{x}^j, \mathbf{w}) + (1 - y^j) \ln P(y^j = 0|\mathbf{x}^j, \mathbf{w}) \\ &= \boxed{\sum_j y^j(w_0 + \sum_i^n w_i x_i^j) - \ln(1 + exp(w_0 + \sum_i^n w_i x_i^j))} \end{aligned}$$

## Maximizing Conditional Log Likelihood

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

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

$$\begin{aligned} l(\mathbf{w}) &\equiv \ln \prod_j P(y^j|\mathbf{x}^j, \mathbf{w}) \\ &= \sum_j y^j(w_0 + \sum_i^n w_i x_i^j) - \ln(1 + exp(w_0 + \sum_i^n w_i x_i^j)) \end{aligned}$$

Bad news: no closed-form solution to maximize *l*(**w**)

Good news: *l*(**w**) is concave function of **w**!

No local minima

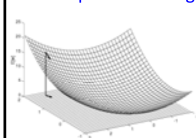
Concave functions easy to optimize



## Optimizing Concave Functions Gradient Ascent

Conditional likelihood for Logistic Regression is concave !

Find optimum with gradient ascent



**Gradient:** $\nabla_{\mathbf{w}} l(\mathbf{w}) = [\frac{\partial l(\mathbf{w})}{\partial w_0}, \ldots, \frac{\partial l(\mathbf{w})}{\partial w_n}]'$

Learning rate, $\eta > 0$

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

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

Gradient ascent is simplest of optimization approaches

e.g., Conjugate gradient ascent much better (see reading)