# Logistic Regression,
# Generative and Discriminative Classifiers

Recommended reading:

• Ng and Jordan paper "On Discriminative vs. Generative classifiers: A comparison of logistic regression and naïve Bayes," A. Ng and M. Jordan, NIPS 2002.

Machine Learning 10-701

Tom M. Mitchell
Carnegie Mellon University

Thanks to Ziv Bar-Joseph, Andrew Moore for some slides

# Overview

Last lecture:

- Naïve Bayes classifier
- Number of parameters to estimate
- Conditional independence

This lecture:

- Logistic regression
- Generative and discriminative classifiers
- (if time) Bias and variance in learning

# Naive Bayes Algorithm

Naive_Bayes_Learn($examples$)

For each target value $v_j$

$\hat{P}(v_j) \leftarrow$ estimate $P(v_j)$

For each attribute value $a_i$ of each attribute $a$

$\hat{P}(a_i|v_j) \leftarrow$ estimate $P(a_i|v_j)$

Classify_New_Instance($x$)

$$v_{NB} = \operatorname*{argmax}_{v_j \in V} \hat{P}(v_j) \prod_{a_i \in x} \hat{P}(a_i|v_j)$$

# Generative vs. Discriminative Classifiers

Training classifiers involves estimating f: X → Y, or P(Y|X)

Generative classifiers:

- Assume some functional form for P(X|Y), P(X)
- Estimate parameters of P(X|Y), P(X) directly from training data
- Use Bayes rule to calculate P(Y|X= $x_i$)

Discriminative classifiers:

1. Assume some functional form for P(Y|X)
2. Estimate parameters of P(Y|X) directly from training data

- Consider learning f: X → Y, where
  - X is a vector of real-valued features, $< X_1 \ldots X_n >$
  - Y is boolean
- So we use a Gaussian Naïve Bayes classifier
  - assume all $X_i$ are conditionally independent given Y
  - model $P(X_i \mid Y = y_k)$ as Gaussian $N(\mu_{ik},\sigma)$
  - model P(Y) as binomial (p)

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

- Consider learning f: X → Y, where
  - X is a vector of real-valued features, < $X_1$ … $X_n$ >
  - Y is boolean
  - assume all $X_i$ are conditionally independent given Y
  - model $P(X_i \mid Y = y_k)$ as Gaussian $N(\mu_{ik},\sigma)$
  - model P(Y) as binomial (p)

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

$$P(Y = 1|X =< x_1, ...x_n >) = \frac{1}{1 + exp(w_0 + \sum_i w_i x_i)}$$

# Logistic regression

- Logistic regression represents the probability of category *i* using a linear function of the input variables:

$$P(Y=i \mid X=x)=g(w_{i0}+w_{i1}x_1+\ldots+w_{id}x_d)$$

where for *i*<*k*

$$g(z_i)=\frac{e^{z_i}}{1+\sum_{j=1}^{K-1} e^{z_j}}$$

and for *k*

$$g(z_k)=\frac{1}{1+\sum_{j=1}^{K-1} e^{z_j}}$$

# Logistic regression

- The name comes from the **logit** transformation:

$$\log \frac{p(Y = i \mid X = x)}{p(Y = K \mid X = x)} = \log \frac{g(z_i)}{g(z_k)} = w_0 + w_{i1} x_1 + \ldots + w_{id} x_d$$

# Binary logistic regression

- We only need one set of parameters

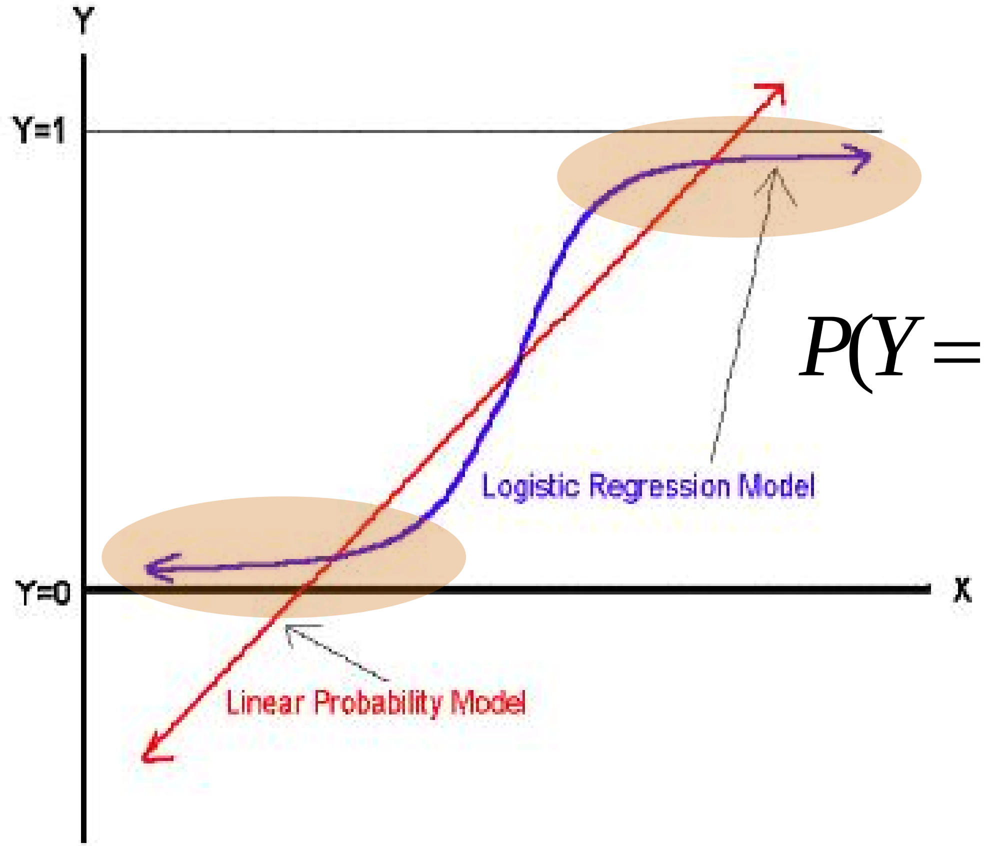
$$p(Y=1|X=x) = \frac{e^{w_0+w_1x_1+\ldots+w_dx_d}}{1+e^{w_0+w_1x_1+\ldots+w_dx_d}}$$

$$= \frac{1}{1+e^{-(w_0+w_1x_1+\ldots+w_dx_d)}}$$

$$= \frac{1}{1+e^{-z}}$$

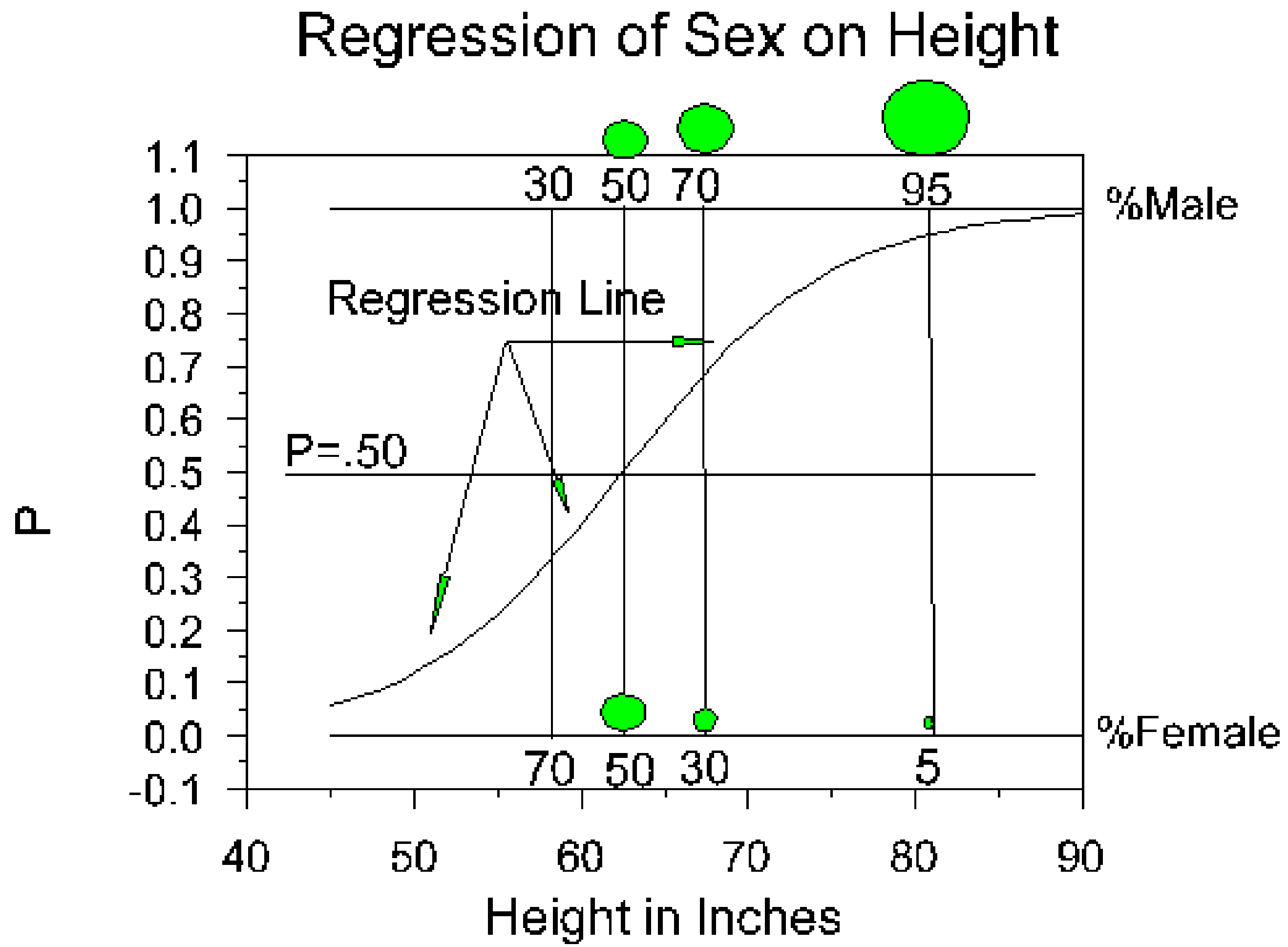
- This results in a "squashing function" which turns linear predictions into probabilities

# Logistic regression vs. Linear regression




$$P(Y=1|X=x) = \frac{1}{1+e^{-z}}$$

# Example



Regression of Sex on Height
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
P
30 50 70
95
%Male
Regression Line
P=.50
%Female
70 50 30
5
40
50
60
70
80
90
Height in Inches

# Log likelihood

$$l(w) = \sum_{i=1}^{N} y_i \log p(x_i; w) + (1 - y_i) \log(1 - p(x_i; w))$$

# Log likelihood

$$l(w) = \sum_{i=1}^{N} y_i \log p(x_i; w) + (1 - y_i) \log(1 - p(x_i; w))$$

$$= \sum_{i=1}^{N} y_i \log \frac{p(x_i; w)}{(1 - p(x_i; w))} + \log(\frac{1}{1 + e^{x_i w}})$$

$$= \sum_{i=1}^{N} y_i x_i w - \log(1 + e^{x_i w})$$

- Note: this likelihood is a concave in *w*

# Maximum likelihood estimation

$$\frac{\partial}{\partial w_j} l(w) = \frac{\partial}{\partial w_j} \sum_{i=1}^{N} \{ y_i x_i w - \log(1 + e^{x_i w}) \}$$

$$= \ldots$$

$$= \sum_{i=1}^{N} x_{ij} \underbrace{(y_i - p(x_i, w))}_{\text{prediction error}}$$

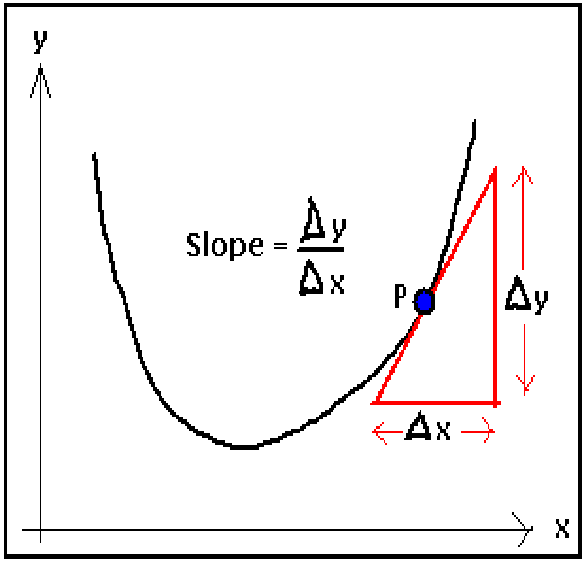
Common (but not only) approaches:

Numerical Solutions:

- Line Search
- Simulated Annealing
- Gradient Descent
- Newton's Method
- Matlab glmfit function

**No close form solution!**

# Gradient descent



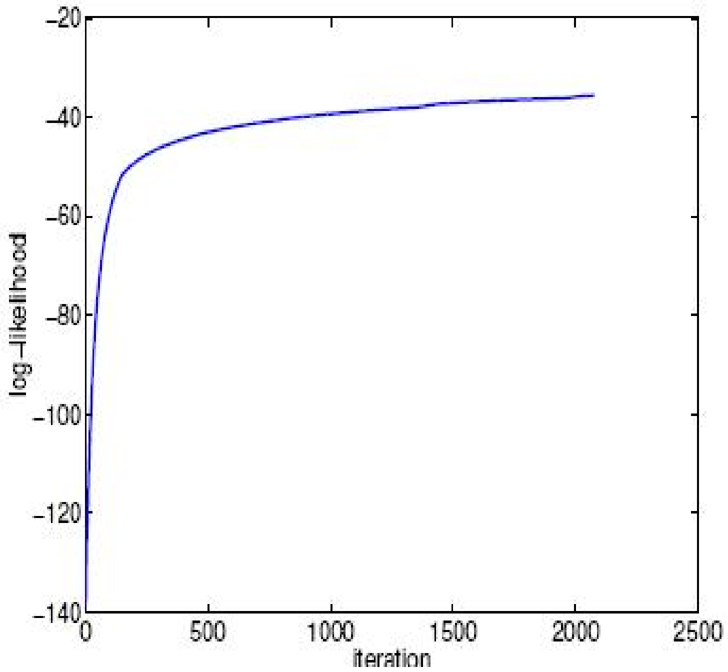
# Gradient ascent

$$w_j^{t+1} \leftarrow w_j^t + \varepsilon \sum_i (x_{ij}(y_i - p(x_i, w)))$$

- Iteratively updating the weights in this fashion increases likelihood each round.
- We eventually reach the maximum
- We are near the maximum when changes in the weights are small.
- Thus, we can stop when the sum of the absolute values of the weight differences is less than some small number.

# Example

- We get a monotonically increasing log likelihood of the training labels as a function of the iterations

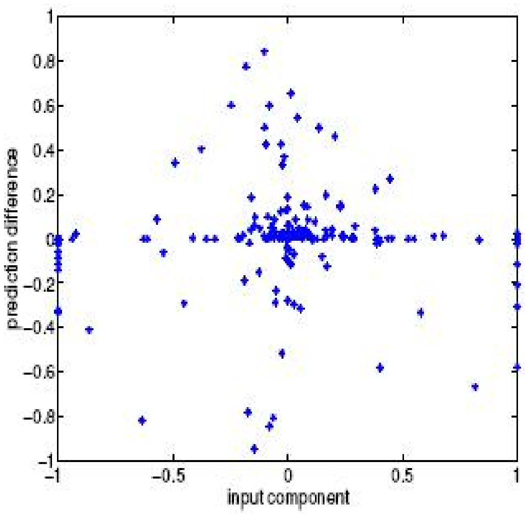


# Convergence

- The gradient ascent learning method converges when there is no incentive to move the parameters in any particular direction:

$$\sum_i (x_{ij}(y_i - p(x_i, w)) = 0 \quad \forall k$$

- This condition means that the prediction error is uncorrelated with the components of the input vector



# Naïve Bayes vs. Logistic Regression

[Ng & Jordan, 2002]

- Generative and Discriminative classifiers
- Asymptotic comparison (# training examples → infinity)
  - when model correct
  - when model incorrect
- Non-asymptotic analysis
  - convergence rate of parameter estimates
  - convergence rate of expected error
- Experimental results

# Generative-Discriminative Pairs

Example: assume Y boolean, X = <$X_1$, $X_2$, …, $X_n$>, where $x_i$ are boolean, perhaps dependent on Y, conditionally independent given Y

Generative model: naïve Bayes:

$$\hat{p}(x_i = 1 | y = b) = \frac{s\{x_i = 1, y = b\} + l}{s\{y = b\} + 2l}$$

*s indicates size of set.*

*l is smoothing parameter*

$$\hat{p}(y = b) = \frac{s\{y = b\}}{\sum_j s\{y = j\}}$$

Classify new example *x* based on ratio

$$\frac{\hat{p}(y = T | x)}{\hat{p}(y = F | x)} = \frac{\hat{p}(y = T) \prod_{i=1}^{n} \hat{p}(x_i | y = T)}{\hat{p}(y = F) \prod_{i=1}^{n} \hat{p}(x_i | y = F)}$$

Equivalently, based on sign of log of this ratio

# Generative-Discriminative Pairs

Example: assume Y boolean, X = <$x_1$, $x_2$, …, $x_n$>, where $x_i$ are boolean, perhaps dependent on Y, conditionally independent given Y

Generative model: naïve Bayes:

$$\widehat{p}(x_i = 1|y = b) = \frac{s\{x_i = 1, y = b\} + l}{s\{y = b\} + 2l}$$

$$\widehat{p}(y = b) = \frac{s\{y = b\}}{\sum_j s\{y = j\}}$$

Classify new example *x* based on ratio

$$\frac{\widehat{p}(y = T|x)}{\widehat{p}(y = F|x)} = \frac{\widehat{p}(y = T)\prod_{i=1}^{n}\widehat{p}(x_i|y = T)}{\widehat{p}(y = F)\prod_{i=1}^{n}\widehat{p}(x_i|y = F)}$$

Discriminative model: logistic regression

$$\widehat{p}(y = T|x; \beta, \theta) = 1/(1 + exp(-\sum_{i=1}^{n}\beta_i x_i - \theta))$$

Note both learn linear decision surface over X in this case

# What is the difference asymptotically?

Notation: let $\epsilon(h_{A,m})$ denote error of hypothesis learned via algorithm A, from *m* examples

- If assumed model correct (e.g., naïve Bayes model), and finite number of parameters, then

$$\epsilon(h_{Dis,\infty}) = \epsilon(h_{Gen,\infty})$$

- If assumed model incorrect

$$\epsilon(h_{Dis,\infty}) \leq \epsilon(h_{Gen,\infty})$$

Note assumed discriminative model can be correct even when generative model incorrect, but not vice versa

# Rate of covergence: logistic regression

Let $h_{Dis,m}$ be logistic regression trained on *m* examples in *n* dimensions. Then with high probability:

$$\epsilon(h_{Dis,m}) \leq \epsilon(h_{Dis,\infty}) + O(\sqrt{\frac{n}{m} \log \frac{m}{n}})$$

Implication: if we want $\epsilon(h_{Dis,m}) \leq \epsilon(h_{Dis,\infty}) + \epsilon_0$

for some constant $\epsilon_0$, it suffices to pick $m = \Omega(n)$

→ Convergences to its classifier, in order of *n* examples

(result follows from Vapnik's structural risk bound, plus fact that VCDim of *n* dimensional linear separators is *n* )

# Rate of covergence: naïve Bayes

Consider first how quickly parameter estimates converge toward their asymptotic values.

Then we'll ask how this influences rate of convergence toward asymptotic classification error.

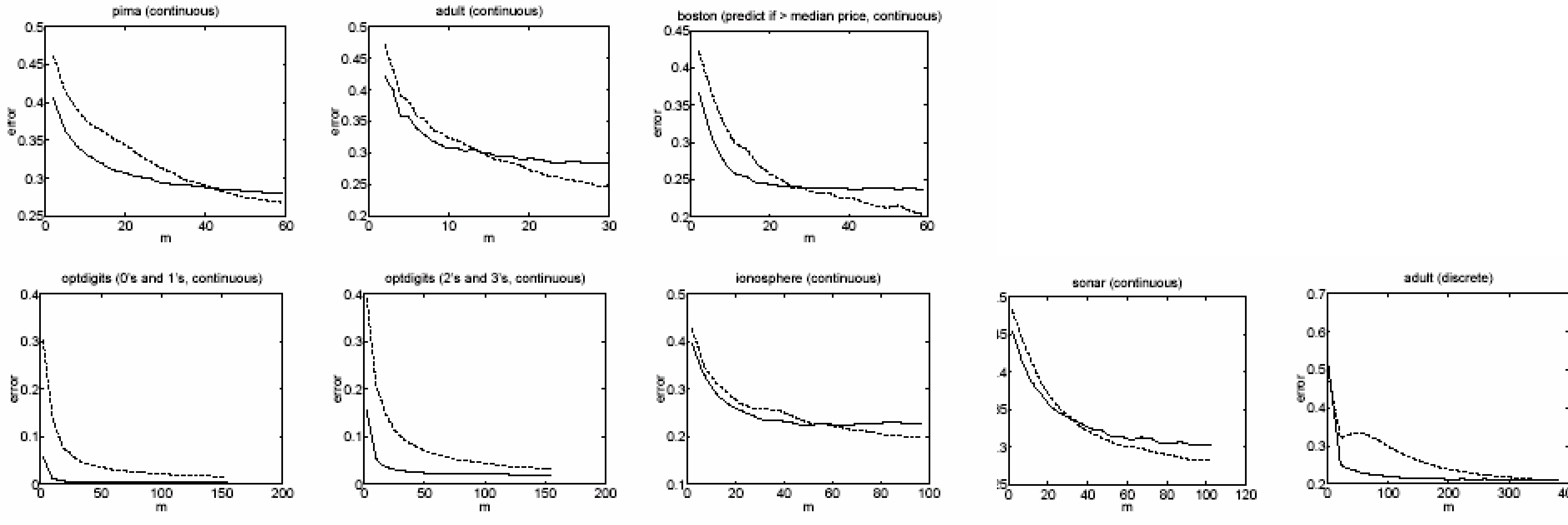
# Rate of covergence: naïve Bayes parameters

Let any $\epsilon_1, \delta > 0$ and any $l \geq 0$ be fixed. Assume that for some fixed $\rho_0 > 0$, we have that $\rho_0 \leq p(y = T) \leq 1 - \rho_0$. Let $m = O((1/\epsilon_1^2)\log(n/\delta))$. Then with probability at least $1 - \delta$, after $m$ examples:

1. For discrete inputs, $|\widehat{p}(x_i|y = b) - p(x_i|y = b)| \leq \epsilon_1$, and $|\widehat{p}(y = b) - p(y = b)| \leq \epsilon_1$, for all i, b.

2. For continuous inputs, $|\widehat{\mu}_{i|y=b} - \mu_{i|y=b}| \leq \epsilon_1$, and $|\widehat{\sigma}_i^2 - \sigma_i^2| \leq \epsilon_1$, for all i, b.

# Some experiments from UCI data sets



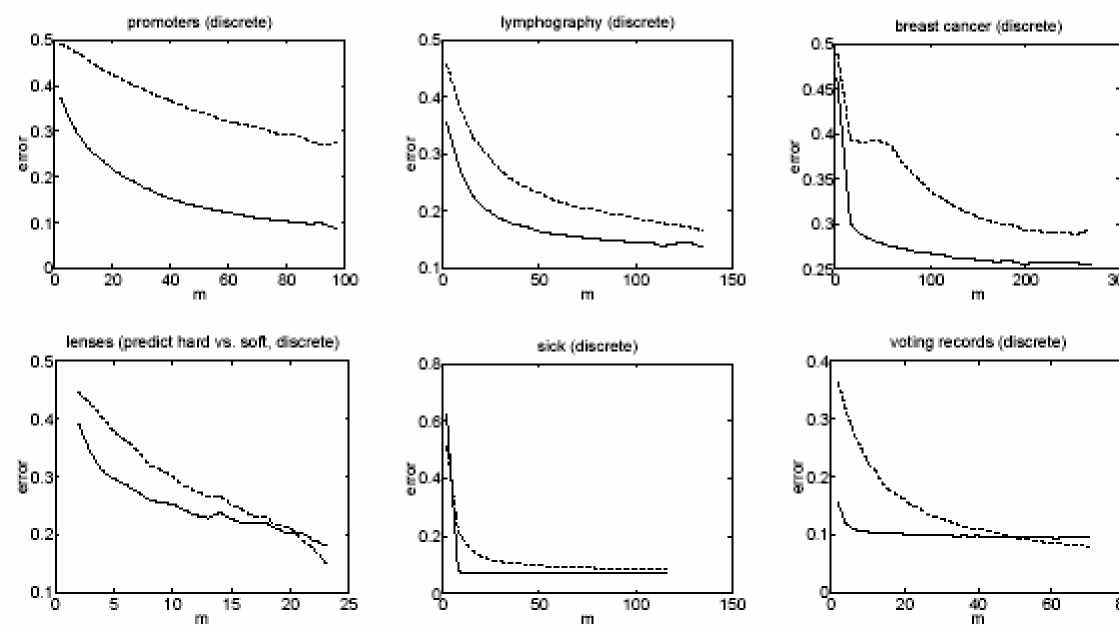




Figure 1: Results of 15 experiments on datasets from the UCI Machine Learnin
repository. Plots are of generalization error vs. $m$ (averaged over 1000 rando
train/test splits). Dashed line is logistic regression; solid line is naive Bayes.

# What you should know:

- Logistic regression
  - What it is
  - How to solve it
  - Log linear models
- Generative and Discriminative classifiers
  - Relation between Naïve Bayes and logistic regression
  - Which do we prefer, when?
- Bias and variance in learning algorithms

# Acknowledgment

Some of these slides are based in part on slides from previous machine learning classes taught by Ziv Bar-Joseph, Andrew Moore at CMU, and by Tommi Jaakkola at MIT.

I thank them for providing use of their slides.