

On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes

Andrew Ng and Michael Jordan
NIPS, 2002

CSE 254: Seminar on Learning Algorithms
Professor: Charles Elkan
Student: Aldebaro Klautau

April 11, 2002

1

Motivation

- **Generative** classifiers are **simpler** to train: estimate $p(x,y)$
- **Discriminative** classifiers can achieve **better accuracy**: "directly" estimate $p(y/x)$
- First reaction: buy a modern computer and use discriminative classifiers
- But, are there cases where generative classifiers outperform discriminative ones ?

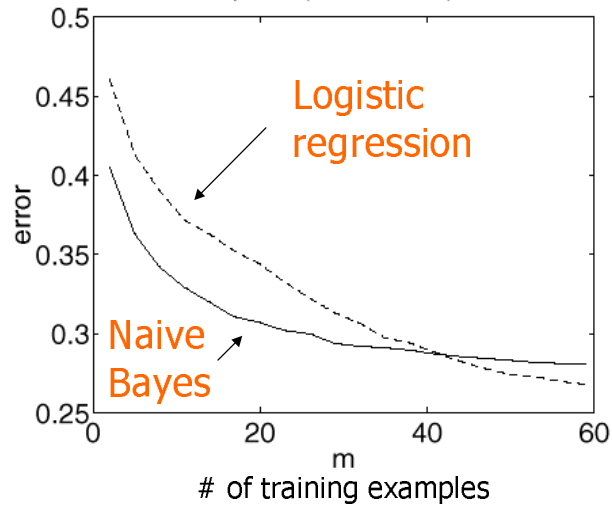


2



There are two regimes

Dataset pima (continuous)



3

Outline

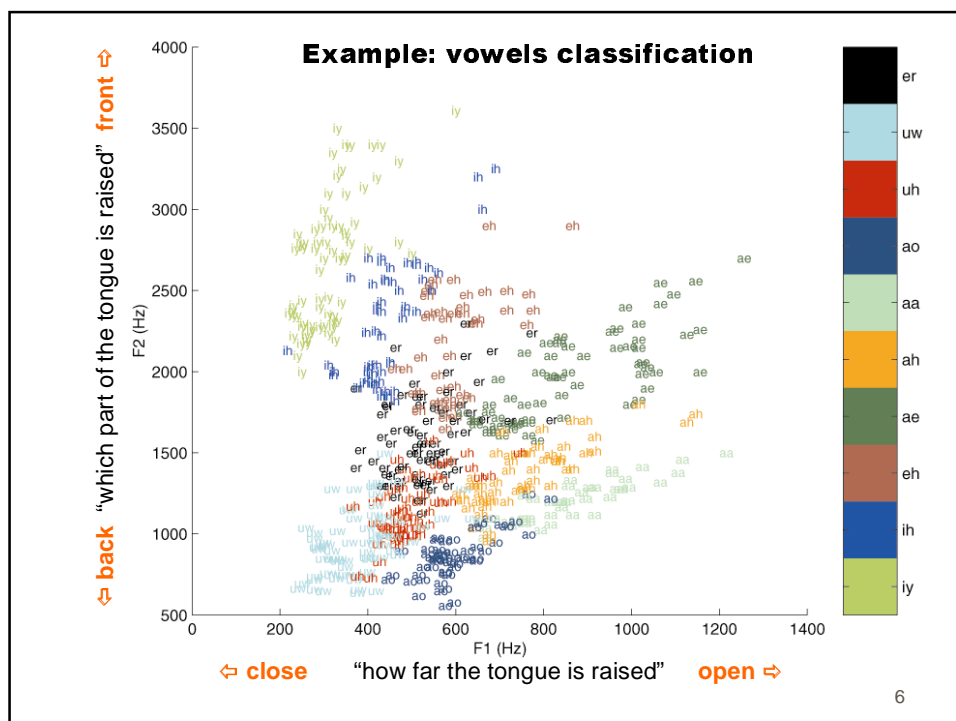
- Supervised classification problem
- Discriminative and generative classifiers
- Naive Bayes versus logistic regression
- Theoretical analysis
 - Useful bounds
 - Sample averages are close to population means (Lemma 3)
 - Naive Bayes can learn with few training examples (Theorem 4)
- Experimental results
- Conclusions

4

Supervised classification problems

- Example (x,y) is composed by **instance** $x \in \mathfrak{X}$ of **dimension** d and **label** $y \in \{1, 2, \dots, K\}$, where K is the **number of classes**
- The **training set** $\{(x^{(1)}, y^{(1)}), \dots, (x^{(m)}, y^{(m)})\}$ with m iid examples is obtained from a joint **distribution** $D = p(x,y)$
- **Task**: given x , identify correct label
- **Goal**: find classifier h that minimizes the **generalization error** $\xi(h) = \Pr_{(x,y) \sim D}[h(x) \neq y]$

5



	Generative	Discriminative
Objective function	joint distribution $p(x,y)$	Conditional distribution $p(y/x)$
Parametric assumptions	Class densities	Class boundaries
Parameter estimation	"easy"	"hard"
Advantages	More efficient if model correct, borrows strength from priors	More flexible, robust because of fewer assumptions
Disadvantages	Bias if model is incorrect	May also be biased. Ignores information from priors

Adapted from "Discriminative vs informative classifiers"
Y. Rubinstein and T. Hastie, KDD'97

11

Naive Bayes classifier

- Assume feature values are independent given the label value:

$$p(\bar{x} / y) = p(x_1 / y)p(x_2 / y)\dots p(x_d / y)$$
- Therefore,

$$L_{\text{Gen}}(\bar{x}) = \log\left(\frac{\hat{p}(\bar{x} / y = T)}{\hat{p}(\bar{x} / y = F)}\right) + \log\left(\frac{\hat{p}(y = T)}{\hat{p}(y = F)}\right)$$
- becomes

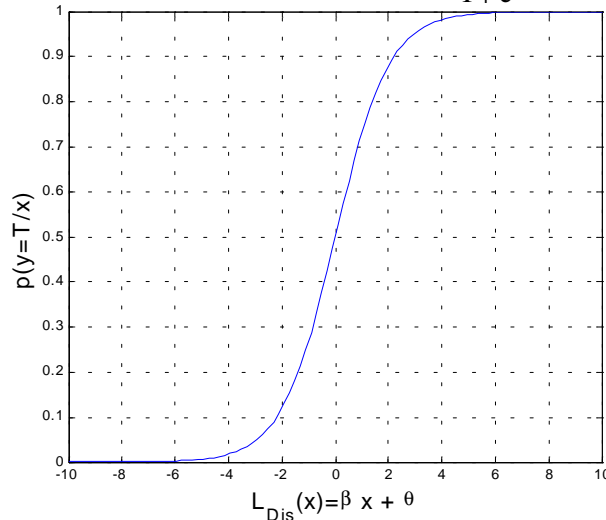
$$L_{\text{Gen}}(\bar{x}) = \sum_{i=1}^d \log\left(\frac{\hat{p}(x_i / y = T)}{\hat{p}(x_i / y = F)}\right) + \log\left(\frac{\hat{p}(y = T)}{\hat{p}(y = F)}\right)$$

12

Logistic regression

- Parameters: coefficients $\beta \in \mathcal{R}^d$ and threshold θ

- Assumption: $p(y = T/x) = \frac{1}{1 + e^{-\beta^T x - \theta}} = \frac{1}{1 + e^{-L_{\text{Dis}}(x)}}$



Note:

$$p(y = T/x) = \frac{e^{L_{\text{Dis}}(x)}}{1 + e^{L_{\text{Dis}}(x)}}$$

$$p(y = F/x) = \frac{1}{1 + e^{L_{\text{Dis}}(x)}}$$

15

Logistic regression (cont.)

- Trained using maximum likelihood estimation

$$(\hat{\beta}, \hat{\theta})_{\text{ML}} = \arg \max_{\beta, \theta} \prod_{j=1}^m \hat{p}(y^{(j)} / x^{(j)}; \beta, \theta)$$

- Assuming labels F and T are 0 and 1, respectively

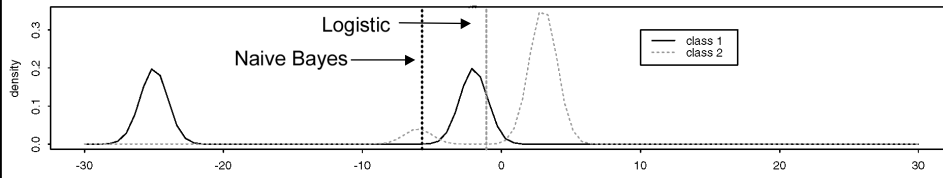
$$(\hat{\beta}, \hat{\theta})_{\text{ML}} = \arg \max_{\beta, \theta} \prod_{j=1}^m \frac{e^{(\beta^T x^{(j)} + \theta)y^{(j)}}}{1 + e^{(\beta^T x^{(j)} + \theta)}}$$

- No closed-form solution. Use Newton methods

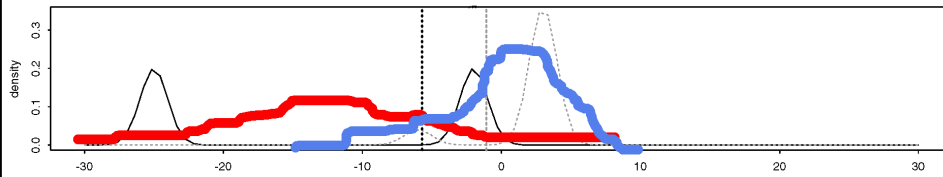
16

A generative method is worse if the generative model is wrong

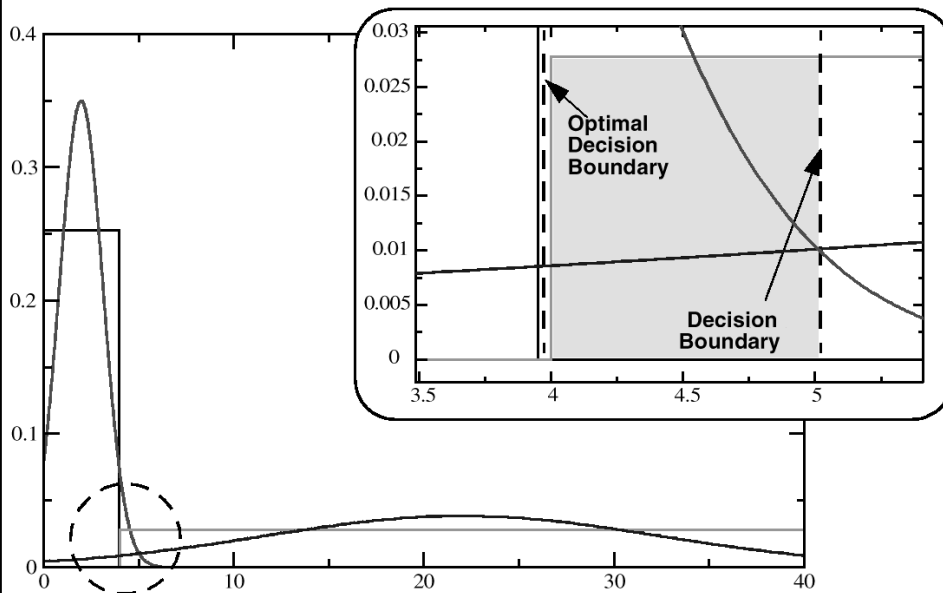
Class densities and decision boundaries



Estimated Gaussians for Naive Bayes



A generative method is worse if the generative model is wrong



Summary of paper contributions

- Theoretical analysis showing that naive Bayes can learn with few training examples
- Outline of the analysis
 - Empirical averages obtained with $m = O(\log d)$ training examples are close to the corresponding population means
 - The Naive Bayes classification rule is a simple function of empirical averages
 - Therefore, the decisions based on empirical averages are close to the decisions derived from infinite training data
- Experimental results

19

Theoretical analysis: Assumptions

- Interested only on generative-discriminative *pairs* $h_{\text{Gen}}, h_{\text{Dis}}$ from **same** parametric **family** of models
- Let $h_{\text{Gen},\infty}, h_{\text{Dis},\infty}$ be their **asymptotic** versions (obtained with infinite amount of data)
- Behavior of h_{Dis} is “fairly well-understood”
- **Proposition 1:** Let $h_{\text{Gen}}, h_{\text{Dis}}$ be any *pair*, then

$$\xi(h_{\text{Gen},\infty}) \geq \xi(h_{\text{Dis},\infty})$$

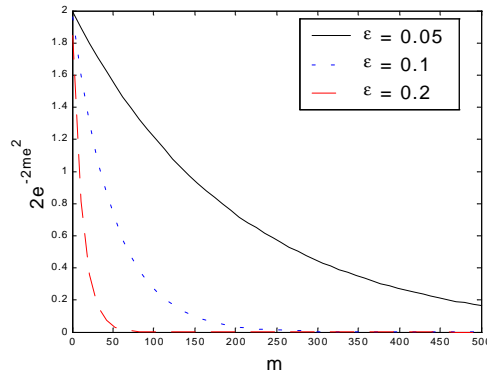
- **Proposition 2:** Let h_{Dis} be logistic regression, then with high probability

$$\xi(h_{\text{Dis}}) \leq \xi(h_{\text{Dis},\infty}) + O\left(\sqrt{\frac{d}{m} \log \frac{m}{d}}\right)$$

20

**Useful bounds:
Hoeffding (or Chernoff) inequality**

- Assume a biased coin with probability of "1" equal to p .
- Tossing it m times gives estimate \hat{p}
- It can be shown* that $\Pr(|\hat{p} - p| \geq \epsilon) \leq 2e^{-2m\epsilon^2}$



* e.g., page 19 of "Neural network learning: Theoretical foundations", by M. Anthony and P. Bartlett, 1999.

**Useful bounds (cont.):
Union bound (or Boole inequality)**

- Assume two events A and B, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \leq P(A) + P(B)$$
- Example: a binary string x of length d is transmitted over a noisy channel. Let δ_i be the probability of flipping the i -th bit (error) and δ_T the probability of receiving a string different than x , then

$$\delta_T \leq \sum_i^d \delta_i$$

- Obs: bound is used to prove Lemma 3

$$\delta_T \leq \delta_1 + 2d\delta_2 \leq \delta$$

Sample averages are close to population means (Lemma 3)

- Let any $\varepsilon_1, \delta > 0$ be fixed. Assume that for some fixed $\rho_0 > 0$, we have that $\rho_0 \leq p(y=T) \leq 1-\rho_0$. Let

$$m = O\left(\frac{1}{\varepsilon_1^2} \log\left(\frac{d}{\delta}\right)\right). \text{ Then with probability at least } 1 - \delta:$$

- In the case of discrete inputs, for all features $i=1, \dots, d$ and $b \in \{T, F\}$:

$$|\hat{p}(x_i / y = b) - p(x_i / y = b)| \leq \varepsilon_1$$

$$|\hat{p}(y = b) - p(y = b)| \leq \varepsilon_1$$

23

Proof of Lemma 3

- Assume $K = 2$ classes and a binary input alphabet $x_i \in \{0, 1\}$
- The maximum likelihood estimate of priors is

$$\hat{p}(y = b) = \frac{\text{count}(y = b)}{m}$$

- The Hoeffding / Chernoff bound $\Pr(|\hat{p} - p| \geq \varepsilon) \leq 2e^{-2m\varepsilon^2}$ implies

$$\Pr(|\hat{p}(y = b) - p(y = b)| \geq \varepsilon_1) \leq 2e^{-2m\varepsilon_1^2} = \delta_1$$

- Example: Pick $\varepsilon_1 = 0.1$ and $\delta_1 = 0.01$, then

$$m = \frac{1}{2\varepsilon_1^2} \log\left(\frac{2}{\delta_1}\right) \approx 265$$

- If $\varepsilon_1 = 0.01$, $\delta_1 = 0.1$, $m \approx 1,500$

24

Proof of Lemma 3 (cont.)

- The maximum likelihood estimate per feature is

$$\hat{p}(x_i = a / y = b) = \frac{\text{count}(x_i = a \wedge y = b)}{\text{count}(y = b)}$$

- Assume there are at least γm examples per class, the Hoeffding / Chernoff bound gives

$$\Pr(|\hat{p}(x_i / y = b) - p(x_i / y = b)| \geq \epsilon_1) \leq 2e^{-2\gamma m \epsilon_1^2} = \delta_2$$

- Note that $2e^{-2m\epsilon_1^2} \leq 2e^{-2\gamma m \epsilon_1^2}$, because $0 < \gamma < 1/2$, i.e., $\delta_1 < \delta_2$
- Since $b \in \{T, F\}$, there are $2d$ such probabilities and the overall chance of error, by the Union bound, is at most $\delta_1 + 2d\delta_2$

25

Proof of Lemma 3 (cont.)

- Want to guarantee

$$\delta \geq \delta_2 + 2d\delta_2 \geq \delta_1 + 2d\delta_2$$

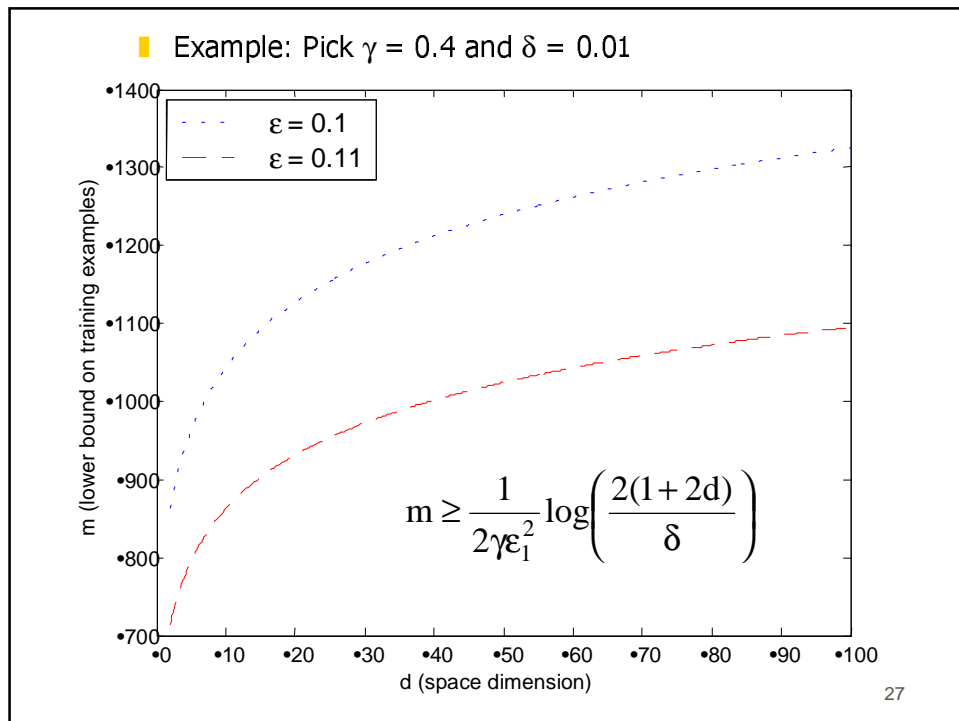
- or equivalently

$$\delta \geq 2e^{-2\gamma m \epsilon_1^2} (1 + 2d)$$

- Therefore

$$m \geq \frac{1}{2\gamma\epsilon_1^2} \log\left(\frac{2(1+2d)}{\delta}\right) = O\left(\frac{1}{\epsilon_1^2} \log\left(\frac{d}{\delta}\right)\right) \quad \square$$

26



■ With **high probability**, when the parameters of h_{Gen} are estimated from set with a **number of samples** m that is only **logarithmic** in the dimension d , they are uniformly close to their correspondent values in $h_{\text{Gen},\infty}$

■ Given

$$L_{\text{Gen}}(\bar{x}) = \sum_{i=1}^d \log\left(\frac{\hat{p}(x_i / y = T)}{\hat{p}(x_i / y = F)}\right) + \log\left(\frac{\hat{p}(y = T)}{\hat{p}(y = F)}\right)$$

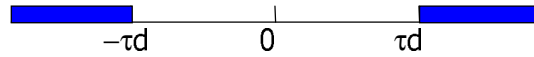
$$L_{\text{Gen},\infty}(\bar{x}) = \sum_{i=1}^d \log\left(\frac{p(x_i / y = T)}{p(x_i / y = F)}\right) + \log\left(\frac{p(y = T)}{p(y = F)}\right)$$

■ What can be said about $\xi(h_{\text{Gen},\infty})$ and $\xi(h_{\text{Gen}})$?

28

Naive Bayes can learn with few $O(\log d)$ training examples (Theorem 4)

- Let τd be a threshold of the score $L_{\text{Gen},\infty}$



- Given an example (x,y) , define the following events:
 - M_T is a "near-miss" $L_{\text{Gen},\infty}(x) \in [0, \tau d]$ when $y = T$
 - M_F is a "near-miss" $L_{\text{Gen},\infty}(x) \in [-\tau d, 0]$ when $y = F$

- Let

$$G(\tau) = \Pr_{(x,y) \sim D}(M_T \vee M_F)$$

- Theorem 4 states that with high probability

$$\xi(h_{\text{Gen}}) \leq \xi(h_{\text{Gen},\infty}) + G\left(O\left(\sqrt{\frac{\log d}{m}}\right)\right)$$

29

Proof of Theorem 4 (sketch)

- Choose m such that

$$m \geq \frac{1}{2\gamma\epsilon_1^2} \log\left(\frac{2(1+2d)}{\delta}\right)$$

- Lemma 3 ensures that with high probability, \hat{p} 's are apart from p 's by at most ϵ_1

$$\epsilon_1 = \sqrt{\frac{1}{2\gamma m} \log\left(\frac{2(1+2d)}{\delta}\right)} = O\left(\sqrt{\frac{\log d}{m}}\right)$$

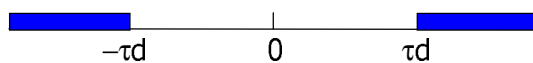
- This implies that

$$|L_{\text{Gen}}(x) - L_{\text{Gen},\infty}(x)| \leq O\left(d \sqrt{\frac{\log d}{m}}\right)$$

30

Proof of Theorem 4 (sketch, cont.)

■ Let $\tau = O\left(\sqrt{\frac{\log d}{m}}\right)$



■ If $|L_{\text{Gen}}(x) - L_{\text{Gen},\infty}(x)| \leq O\left(d \sqrt{\frac{\log d}{m}}\right) = O(\tau d)$

■ the probability of h_{Gen} being wrong while $h_{\text{Gen},\infty}$ is correct is $G(\tau) = \Pr_{(x,y) \sim D}(M_T \vee M_F)$, therefore

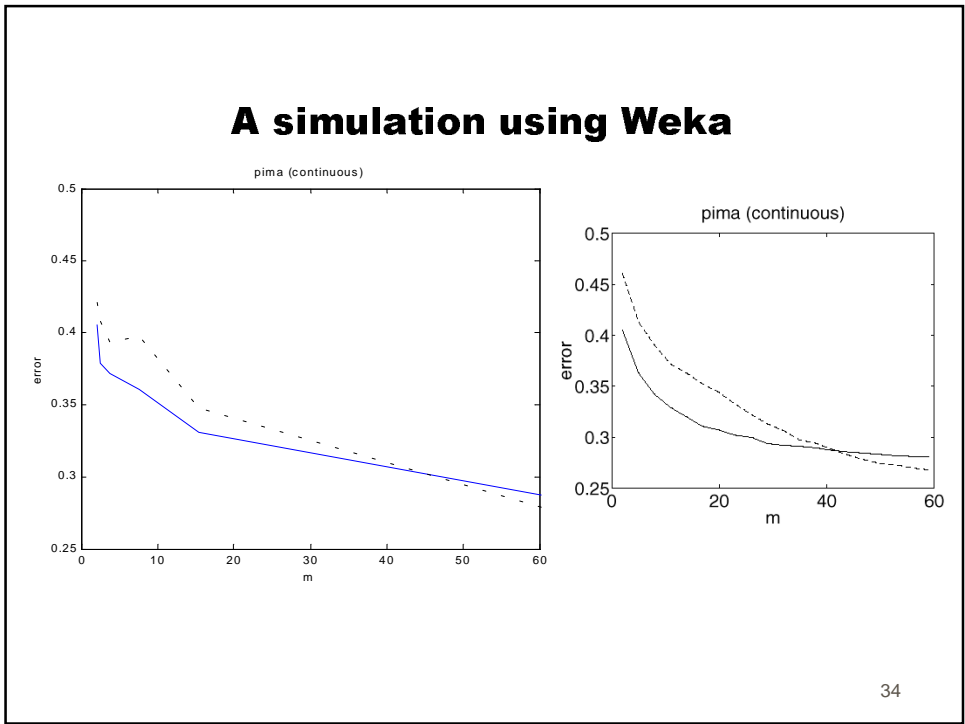
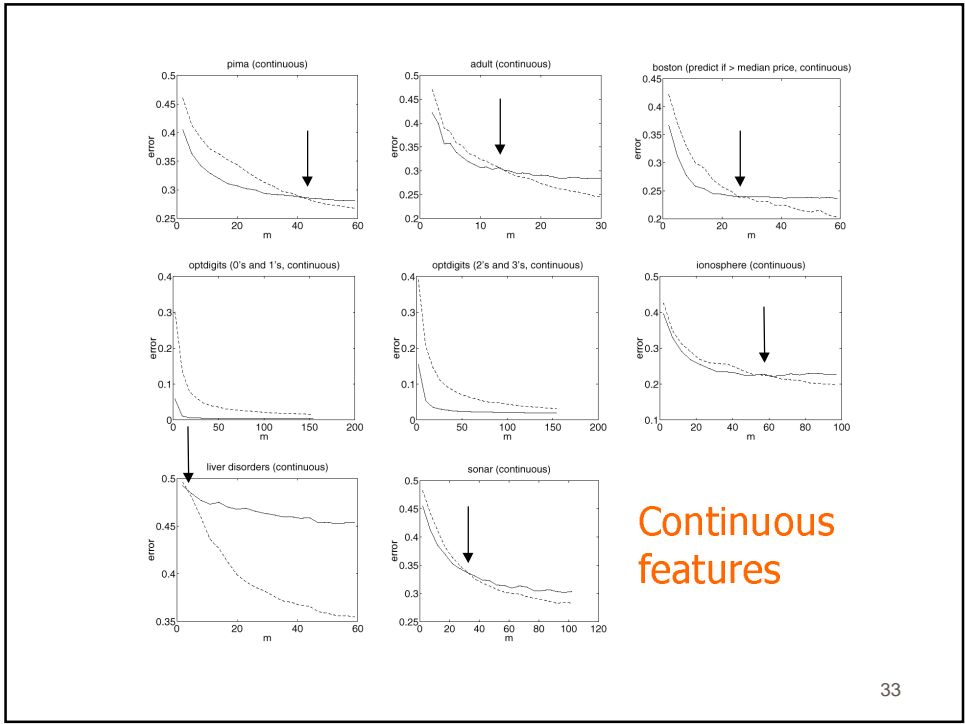
$$\xi(h_{\text{Gen}}) \leq \xi(h_{\text{Gen},\infty}) + G\left(O\left(\sqrt{\frac{\log d}{m}}\right)\right) \quad \square$$

31

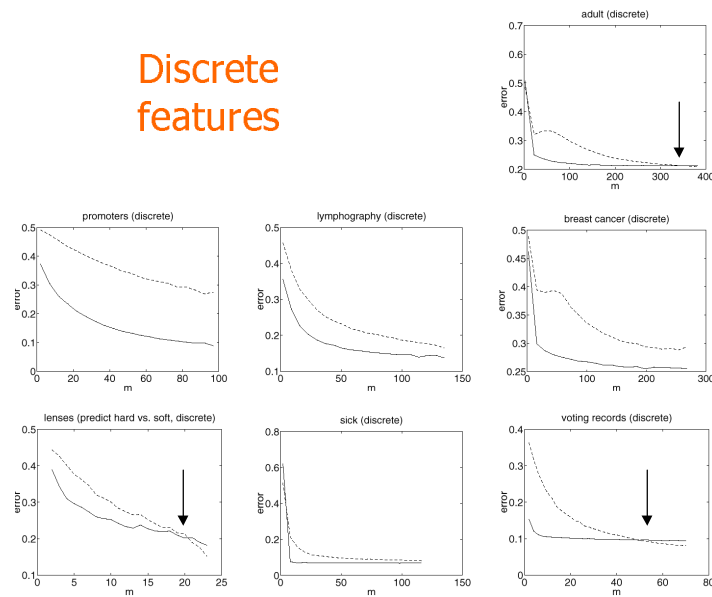
Experimental results

- 15 datasets from UCI repository
 - 8 with only continuous features
 - 7 with only discrete features
- Only binary classification problems
- Did not use any regularization method
 - Potentially hurts logistic regression
- Used Laplace smoothing for naive Bayes

32



Discrete features



35

Conclusions

- Naive Bayes can learn with few $O(\log d)$ training examples
- Claimed existence of two regimes:
 - naive Bayes wins with few data
 - logistic regression wins with much data (lower asymptotic error)
- Experiments confirmed the claim, specially with continuous features

36