Generative vs. Discriminative Classifiers

Logistic Regression

- 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 I Y= y_k) as Gaussian N(μ_{ik} , σ_i)
  - Model P(Y) as Bernoulli (π)
- Then P(Y|X) is of this form, and we can directly estimate W

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

- Furthermore, same holds if the X_i are boolean
  - Trying proving that to yourself
- Train by gradient ascent estimation of w's (no assumptions!)

MLE vs MAP

- Maximum conditional likelihood estimate

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

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

- Maximum a posteriori estimate with prior W~N(0,σI)

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

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

Generative vs. Discriminative Classidiers

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

Generative classifiers (e.g., Naïve Bayes)

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

Discriminative classifiers (e.g., Logistic regression)

- Assume some functional form for P(Y|X)
- Estimate parameters of P(Y|X) directly form training data

- **NOTE!** Even though our derivation of the form of P(Y|X) made GNB-style assumptions, the *training procedure* for Logistic Regression does not !

Use Naïve Bayes or Logistic Regression ?

Consider

- Restrictiveness of modeling assumptions

- Rate of convergence (in amount of training data) toward asymptotic hypothesis
  - i.e., the learning curve

Naïve Bayes vs Logistic Regression

Consider Y boolean, X_i continuous, $X = < X_1 \ldots X_n >$

Number of parameters to estimate :

- NB :

- LR :

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

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

G.Naïve Bayes vs. Logistic Regression

[Ng & Jordan, 2002]

Recall two assumptions deriving form of LR from Gnbayes:

1. X_i conditionally independent of X_k given Y
2. $P(X_i | Y=y_k) = N(\mu_{ik}, \sigma_i)$, ← not $N(\mu_{ik}, \sigma_{ik})$

Consider three learning methods:

- GNB (assumption 1 only)
- GNB2 (assumption 1 and 2)
- LR

Which method works better if we have *infinite* training data, and ...

- Both (1) and (2) are satisfied
- Neither (1) or (2) is satisfied
- (1) is satisfied, but not (2)

G.Naïve Bayes vs. Logistic Regression

[Ng & Jordan, 2002]

What if we have only finite training data ?

They converage at different rates to their asymptotic (∞ data) error

Let $\epsilon_{A,n}$ refer to expected error lof learning algorithm A after n training examples

Let d be the number of features : <X_1...X_d>

$$\in_{GNB,n} \leq \in_{GNB,\infty} + O\left(\sqrt{\frac{\log d}{n}}\right) \qquad \in_{LR,n} \leq \in_{LR,\infty} + O\left(\sqrt{\frac{d}{n}}\right)$$

So, GNB reguires **O(log d)** to convergen, but **LR** requires **O(d)**

Some experiments from UCI data sets

[Ng & Jordan, 2002]

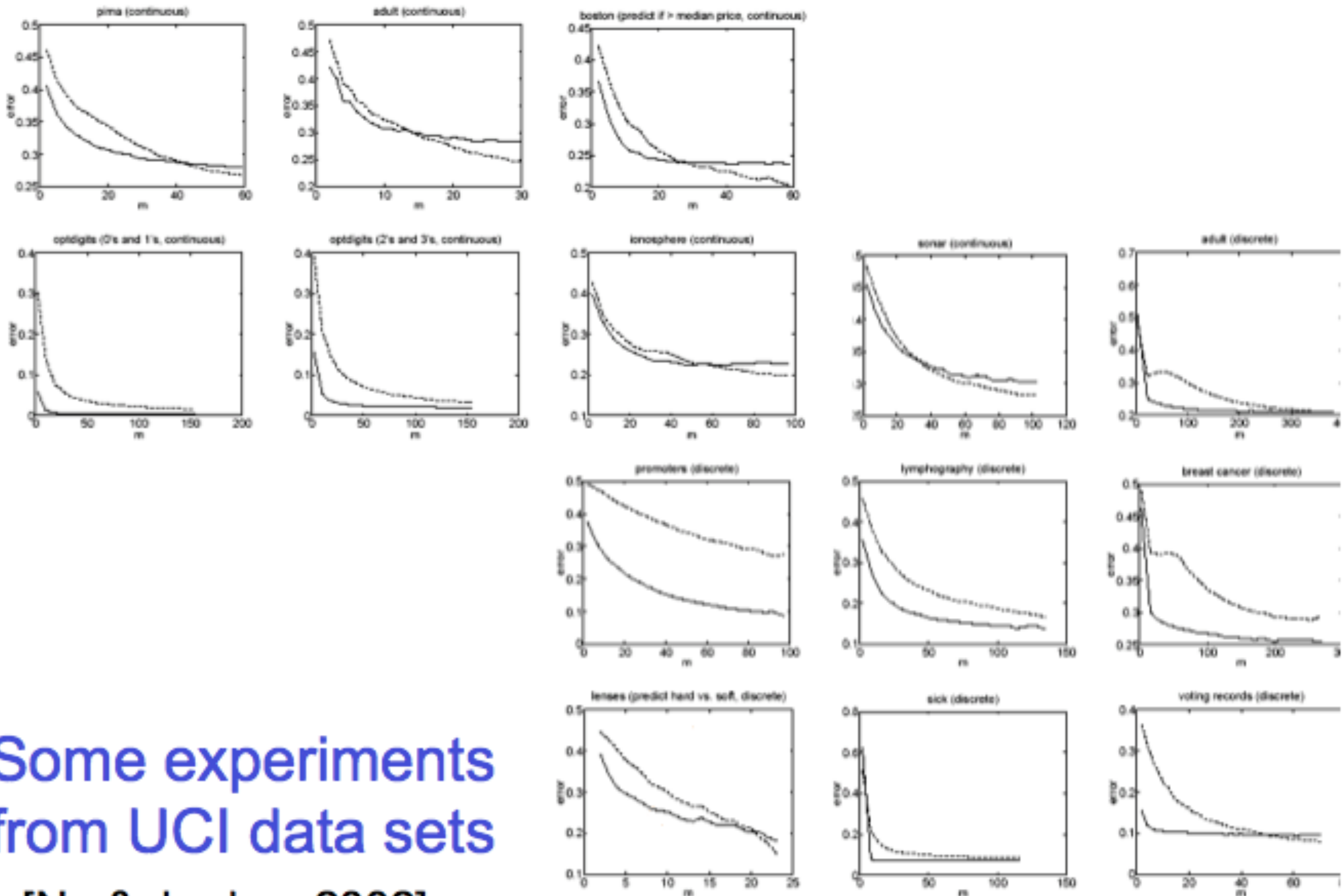

Figure 1: Results of 15 experiments on datasets from the UCI Machine Learnin
repository. Plots are of generalization error vs. m (averaged over 1000 rando
train/test splits). Dashed line is logistic regression; solid line is naive Bayes.

G.Naïve Bayes vs. Logistic Regression

The bottom line :

GNB2 and LR both use linear decision surfaces, GNB need not

Given infinite data, LR is better than GNB2 because *training procedure* does not make assumptions 1 or 2 (though our derivation of the form of P(Y|X) did)

But GNB2 converges more quickly to its perhaps-less-accurate asymptotic error

And GNB is both more bias (assumption1) and less (no assumption2) than LR, so either might beat the other