Logistic Regression

Adrian Barbu

Logistic Regression, 2 Classes

- Naïve Bayes
  - Two classes, Y=0 and Y=1
  - Continuous features $x=(x_1, \ldots, x_M)$
  - Model $P(x_i|Y=y_k)$ as Gaussian $N(\mu_{ik},\sigma_i)$ (variance independent of Y)
  - P(Y) is a 2-bin histogram (Bernoulli), $P(Y=1)=\pi$
- Obtain discriminative classifier:

$$P(Y=1|\mathbf{x}) = \frac{1}{1+\exp(w_0+\sum_i w_i x_i)}$$

$$P(Y=0|\mathbf{x}) = \frac{\exp(w_0+\sum_i w_i x_i)}{1+\exp(w_0+\sum_i w_i x_i)}$$

Derivation

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

$$= \frac{1}{1+\frac{P(Y=0)P(\mathbf{x}|Y=0)}{P(Y=1)P(\mathbf{x}|Y=1)}} = \frac{1}{1+\frac{(1-\pi)}{\pi}\frac{P(\mathbf{x}|Y=0)}{P(\mathbf{x}|Y=1)}}$$

$$= \frac{1}{1+\exp(\ln\frac{(1-\pi)}{\pi}+\ln\frac{P(\mathbf{x}|Y=0)}{P(\mathbf{x}|Y=1)})}$$

$$= \frac{1}{1+\exp(\ln\frac{(1-\pi)}{\pi}+\sum_i \ln\frac{P(x_i|Y=0)}{P(x_i|Y=1)})}$$

$$= \frac{1}{1+\exp(\ln\frac{(1-\pi)}{\pi}+\sum_i \left(\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}x_i+\frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}\right))}$$

$$P(Y=1|\mathbf{x}) = \frac{1}{1+\exp(w_0+\sum_i w_i x_i)}$$

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

Linear Classification

- Thus $$\frac{P(Y=0|\mathbf{x})}{P(Y=1|\mathbf{x})} = \exp(w_0 + \sum_i w_i x_i)$$

  So $$\ln \frac{P(Y=0|\mathbf{x})}{P(Y=1|\mathbf{x})} = w_0 + \sum_i w_i x_i$$

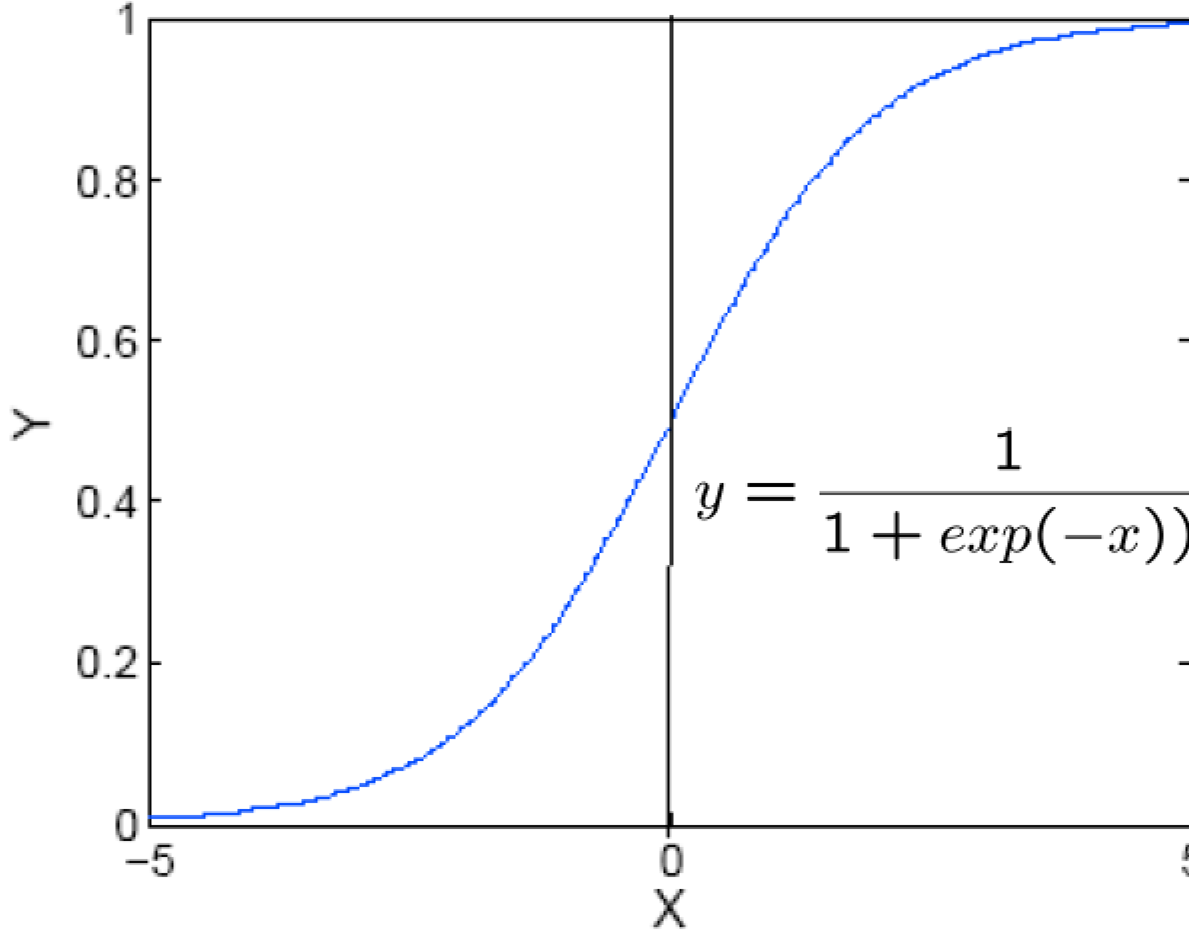
  Say we output Y=0 if $$\frac{P(Y=0|\mathbf{x})}{P(Y=1|\mathbf{x})} > 1$$

  This means

  $$w_0 + \sum_i w_i x_i = \ln \frac{P(Y=0|\mathbf{x})}{P(Y=1|\mathbf{x})} > \ln(1) = 0$$

  i.e. linear classification boundary

Logistic Function



$$y = \frac{1}{1 + exp(-x))}$$

$$x = -(w_0 + \sum_i w_i x_i)$$

Training

- Training data: D={(x^1,Y^1),…,(x^N,Y^N)}
- Find w=(w_1,…,w_M) to maximize the Conditional Likelihood:

$$CL(\mathbf{w}) = \prod_{j=1}^{N} P(Y^j|\mathbf{x}^j, \mathbf{w})$$

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

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

- Take ln

$$L(\mathbf{w}) = \ln \prod_{j=1}^{N} P(Y^j|\mathbf{x}^j, \mathbf{w})$$

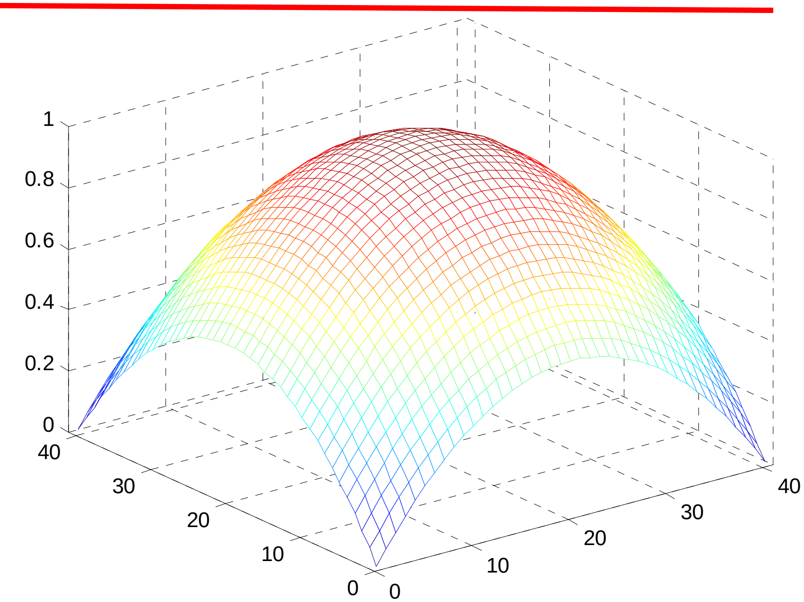
Conditional Likelihood

$$L(\mathbf{w}) = \sum_{j=1}^{N} Y^j \ln P(Y^j = 1|\mathbf{x}^j, \mathbf{w}) + (1 - Y^j) \ln P(Y^j = 0|\mathbf{x}^j, \mathbf{w})$$

$$= \sum_{j=1}^{N} Y^j \ln \frac{P(Y^j = 1|\mathbf{x}^j, \mathbf{w})}{P(Y^j = 0|\mathbf{x}^j, \mathbf{w})} + \ln P(Y^j = 0|\mathbf{x}^j, \mathbf{w})$$

$$= \sum_{j=1}^{N} Y^j (w_0 + \sum_{i=1}^{M} w_i x_i^j) - \ln(1 + \exp(w_0 + \sum_{i=1}^{M} w_i x_i^j))$$

- Good news:
  - L(w) is concave in w
  - One global optimum
- Bad news:
  - No closed form solution



Conditional Likelihood Maximization

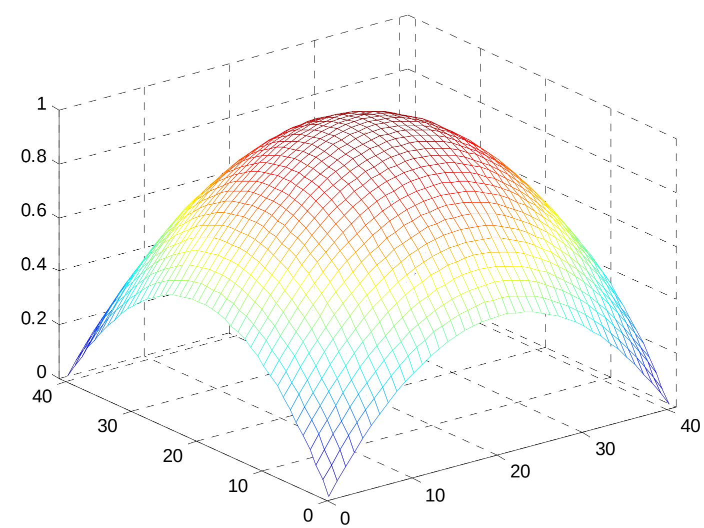
- Gradient of L(w)

$$\frac{\partial L(\mathbf{w})}{\partial w_k} = \sum_{j=1}^{N} x_k^j (Y^j - \frac{\exp(w_0 + \sum_{i=1}^{M} w_i x_i^j)}{1 + \exp(w_0 + \sum_{i=1}^{M} w_i x_i^j)})$$

- Gradient ascent:

x_0^j=1

$$\mathbf{w} \leftarrow \mathbf{w} + \eta \frac{\partial L(\mathbf{w})}{\partial \mathbf{w}}$$



Maximum A Posteriori

- MAP adds a prior on W

$$AP(\mathbf{w}) = P(\mathbf{w}) \prod_{j=1}^{N} P(Y^j|\mathbf{x}^j, \mathbf{w})$$

- Say prior is Gaussian, $N(0,\lambda^{-1}I_{M+1})$
- Obtain gradient ascent equation

$$\mathbf{w} \leftarrow \mathbf{w} - \eta\lambda\mathbf{w} + \eta\frac{\partial L(\mathbf{w})}{\partial \mathbf{w}}$$

- The k-th element update is:

$$w_k \leftarrow w_k - \eta\lambda w_k + \eta \sum_{j=1}^{N} x_k^j (Y^j - \frac{\exp(w_0 + \sum_{i=1}^{M} w_i x_i^j)}{1 + \exp(w_0 + \sum_{i=1}^{M} w_i x_i^j)})$$

Multi-class Logistic Regression

- Assume R classes $Y=\{Y_1,\ldots,Y_R\}$
- Learn R-1 sets of weights
  - For k<R

$$P(Y=Y_k|\mathbf{x}) = \frac{\exp(w_{k0}+\sum_{i=1}^{M} w_{ki}x_i)}{1+\sum_{j=1}^{R-1}\exp(w_{j0}+\sum_{i=1}^{M} w_{ji}x_i)}$$

  - For k=R

$$P(Y=Y_R|\mathbf{x}) = \frac{1}{1+\sum_{j=1}^{R-1}\exp(w_{j0}+\sum_{i=1}^{M} w_{ji}x_i)}$$

- Conditional Log-likelihood not convex anymore
- Training is more complicated

Naïve Bayes vs Logistic Regression

- Number of parameters
  - Naïve Bayes: 4M+1
  - Logistic Regression: M+1

- Parameter estimation
  - Naïve Bayes:
    - Independent since they are not coupled
    - Faster training
  - Logistic Regression:
    - Gradient descent since they are coupled
    - Slower training

Naïve Bayes vs Logistic Regression

Ng & Jordan, 2002

- Asymptotic comparison (# training examples → ∞)

  When model assumptions are
  - Correct: obtain identical classifiers
  - Incorrect: Logistic Regression is less biased, better results
- Non-asymptotic results

  Number of training samples N required:
  - Naïve Bayes: O(log M)
  - Logistic Regression: O(M)
- Conclusion
  - Naïve Bayes: Faster, needs fewer examples but worse results
  - Logistic Regression: Slower, needs more examples but better results

Error Rate

Ng & Jordan, 2002

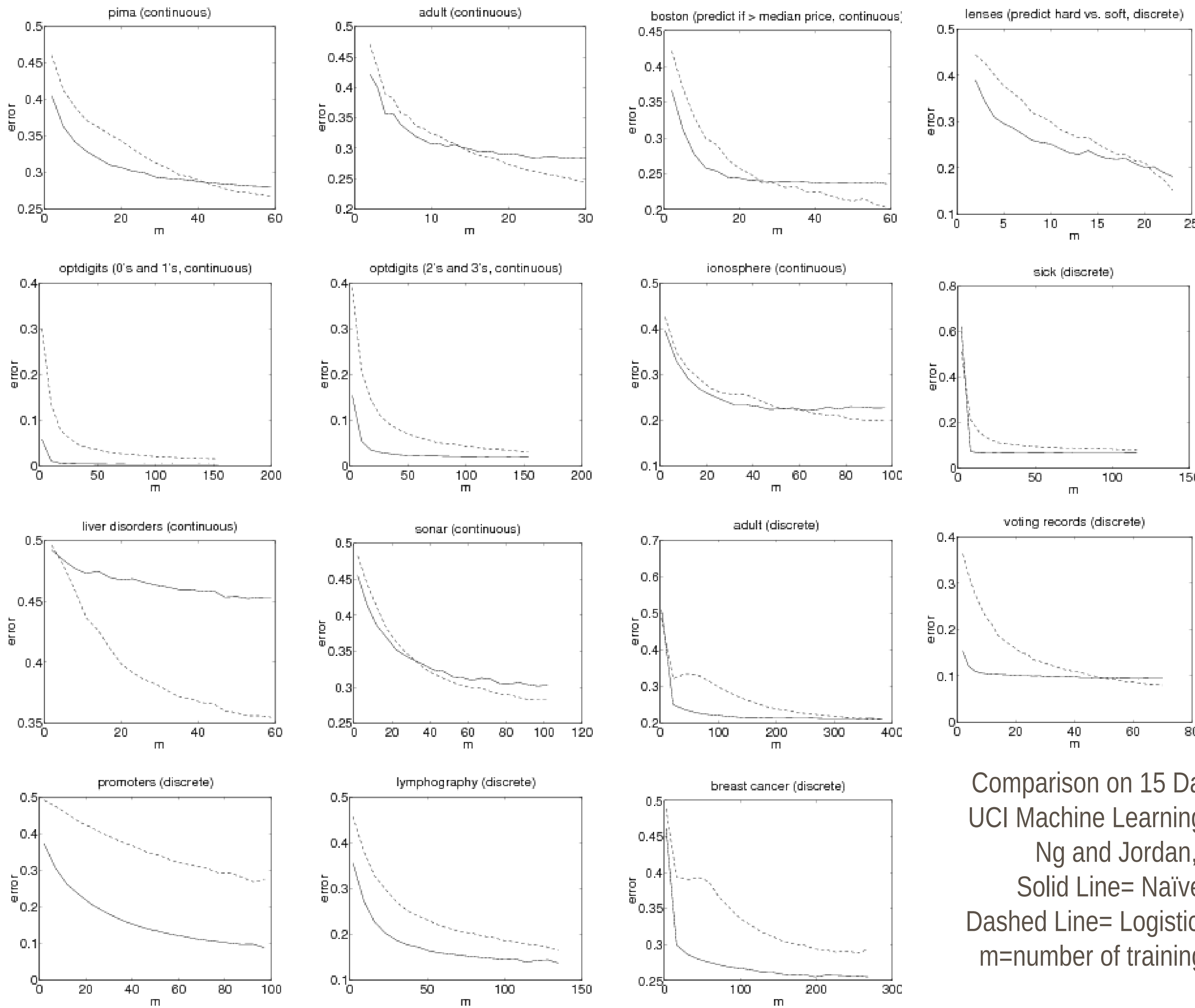
Logistic Regression: M features

- Error for a number of examples N. With high probability

$$err(N) < err(\infty) + O(\sqrt{\frac{M}{N}\log\frac{N}{M}})$$

- So for a good error need about M examples

Naïve Bayes vs Logistic Regression



Comparison on 15 Datasets from
UCI Machine Learning Repository
Ng and Jordan, 2002
Solid Line= Naïve Bayes
Dashed Line= Logistic Regression
m=number of training examples

Conclusions

- Logistic Regression
  - Obtained from Naïve Bayes
  - Discriminative
  - Does not assume conditional independence of features
- Pros
  - Better results than Naïve Bayes
- Cons
  - Slower to train
  - Needs more training examples