Bias and Variance in Learning

Instance-Based Learning

Recommended reading:

- Bias-Variance : Bishop chapter 9.1, 9.2
- Instance-based learning: Mitchell chapter 8.1 – 8.4

Machine Learning 10-701

Tom M. Mitchell
Carnegie Mellon University

Previous lecture:

(see new lecture notes on class website)

- Logistic regression
- Generative and discriminative classifiers
  - E.g, Naïve Bayes and Logistic regression

This lecture:

- Training for logistic regression
- Bias-Variance decomposition of error
- Instance-based learning

Logistic regression

Form of P(Y|X): $$P(Y=0|X) = \frac{1}{1+\exp(w_0 + \sum_{i=1}^{n} w_i X_i)}$$

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

Training: choose weights w_i to maximize

- conditional data likelihood:

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

- or classification accuracy

$$W \leftarrow \arg\max_W \sum_l \delta(Y^l = h(X^l, W))$$

Log likelihood

$$
\begin{aligned}
l(W) &= \sum_l Y^l \ln P(Y^l = 1|X^l, W) + (1 - Y^l) \ln P(Y^l = 0|X^l, W) \\
&= \sum_l Y^l \ln \frac{P(Y^l = 1|X^l, W)}{P(Y^l = 0|X^l, W)} + \ln P(Y^l = 0|X^l, W) \\
&= \sum_l Y^l (w_0 + \sum_i^n w_i X_i^l) - \ln(1 + exp(w_0 + \sum_i^n w_i X_i^l))
\end{aligned}
$$

$$
\frac{\partial l(W)}{\partial w_i} = \sum_l X_i^l (Y^l - \hat{P}(Y^l = 1|X^l, W))
$$

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

Maximizing conditional log likelihood

$$\frac{\partial l(W)}{\partial w_i} = \sum_l X_i^l (Y^l - \hat{P}(Y^l = 1 | X^l, W))$$

Gradient ascent rule:

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

Regularization

The issue: fear of overfitting training data at the expense of poorly fitting future data

The approach: choose weights that maximize a new, *penalized* likelihood function

$$W \leftarrow \arg\max_W \sum_l \ln P(Y^l | X^l, W) - \frac{\lambda}{2} ||W||^2$$

Penalty term,

Regularization term

Regularization

The $||W||^2$ penalty corresponds to adding a Gaussian prior to our weight estimator!

Maximum likelihood estimate:

$$W \leftarrow \arg\max_W P(Data|W)$$

MAP estimate:

$$W \leftarrow \arg\max_W P(W|Data) = \arg\max_W P(Data|W)P(W)$$

$$= \arg\max_W \log P(Data|W) + \log P(W)$$

Regularization

The $||W||^2$ penalty corresponds to adding a Gaussian prior to our weight estimator!

Maximum likelihood:

$$W \leftarrow \arg\max_W P(Data|W)$$

Gaussian N(0,σ)

MAP estimate:

$$W \leftarrow \arg\max_W P(W|Data) = \arg\max_W P(Data|W)P(W)$$

$$= \arg\max_W \log P(Data|W) + \log P(W)$$

c $||W||^2$

Regularization in Logistic Regression

$$\underbrace{W \leftarrow \arg\max_W \sum_l \ln P(Y^l|X^l, W) - \frac{\lambda}{2}||W||^2}_{l(W)}$$

$$\frac{\partial l(W)}{\partial w_i} = \sum_l X_i^l (Y^l - \hat{P}(Y^l = 1|X^l, W)) - \lambda w_i$$

New gradient ascent rule:

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

Generative vs. Discriminative Classifiers

Training classifiers involves estimating f: X $\rightarrow$ 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:

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

G.Naïve Bayes vs. Logistic Regression

[Ng & Jordan, 2002]

- Generative and Discriminative classifiers
- Asymptotic comparison (# training examples → infinity)
  - when model correct
    - GNB, LR produce identical classifiers
  - when model incorrect
    - LR is less biased – does not assume cond indep.
    - therefore expected to outperform GNB

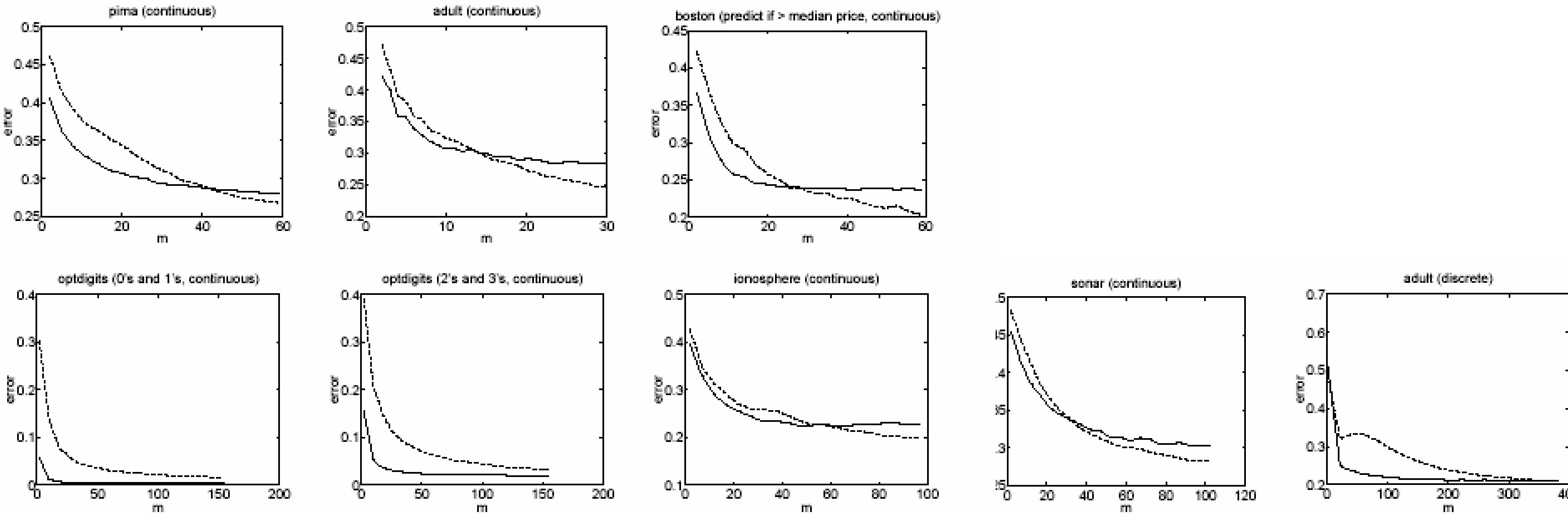
Naïve Bayes vs. Logistic Regression

[Ng & Jordan, 2002]

- Generative and Discriminative classifiers
- Non-asymptotic analysis
  - convergence rate of parameter estimates
    - GNB order log n (# of attributes in X)
    - LR order n

GNB converges more quickly to its (perhaps less helpful) asymptotic estimates

Some experiments from UCI data sets



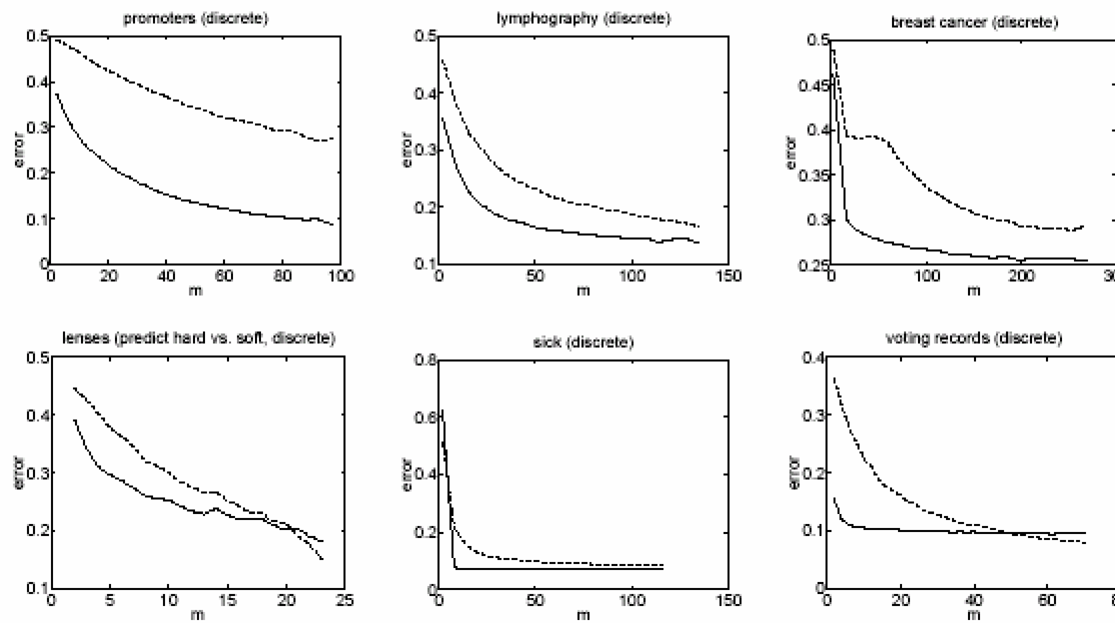


Figure 1: Results of 15 experiments on datasets from the UCI Machine Learnin repository. Plots are of generalization error vs. m (averaged over 1000 randor train/test splits). Dashed line is logistic regression; solid line is naive Bayes.

Bias – Variance decomposition of error

- Consider simple regression problem $f: X \rightarrow Y$

$$y = f(x) = g(x) + \varepsilon$$

noise ~ $N(0,\sigma)$

deterministic

What are sources of prediction error?

Sources of error

- What if we have perfect learner, infinite data?
  - Our learned h(x) satisfies h(x)=g(x)
  - Still have remaining, <u>*unavoidable error*</u> of σ^2 due to noise ε

Sources of error

- What if we have imperfect learner, or only m training examples?
- What is our expected squared error per example
  - Expectation taken over random training sets *D* of size m, drawn from distribution P(X,Y)

$$E_D\left[\int_x \int_y (h(x)-y)^2 p(f(x)=y|x)p(x)dydx\right]$$

Bias-Variance Decomposition of Error

Assume target function: y = f(x) = g(x) + ε

Then expected sq error over fixed size training sets *D* drawn from P(Y,X) can be expressed as sum of three components:

$$E_D\left[\int_x \int_y (h(x) - y)^2 p(y|x)p(x)dydx\right]$$

$$= unavoidableError + bias^2 + variance$$

Where:

$$unavoidableError = \sigma^2$$

$$bias^2 = \int (E_D[h(x)] - g(x))^2 p(x)dx$$

$$variance = \int E_D[(h(x) - E_D[h(x)])^2]p(x)dx$$