CSE 537 Fall 2015

LEARNING FROM EXAMPLES

AIMA CHAPTER 18 (4-5)

Instructor: Sael Lee

Slides are mostly made from AIMA resources,
Andrew W. Moore's tutorials: http://www.cs.cmu.edu/~awm/tutorials and
Bart Selman's Cornell CS4700 decision tree slides

AIMA Chapter 18 (4)

EVALUATING AND CHOOSING THE BEST HYPOTHESIS

CHOOSING BETWEEN HYPOTHESIS

There can be multiple consistent hypothesis

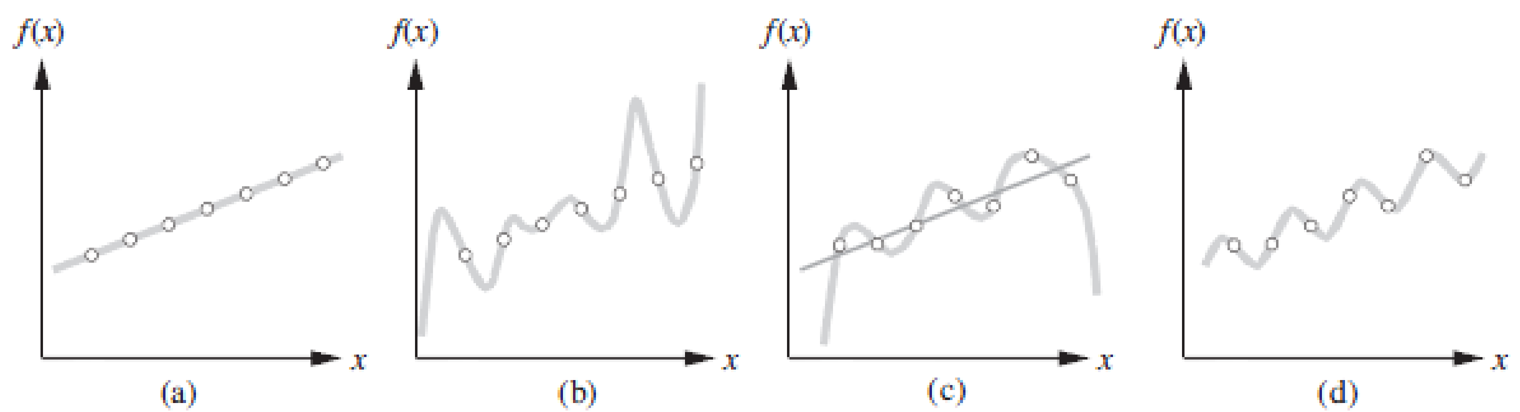
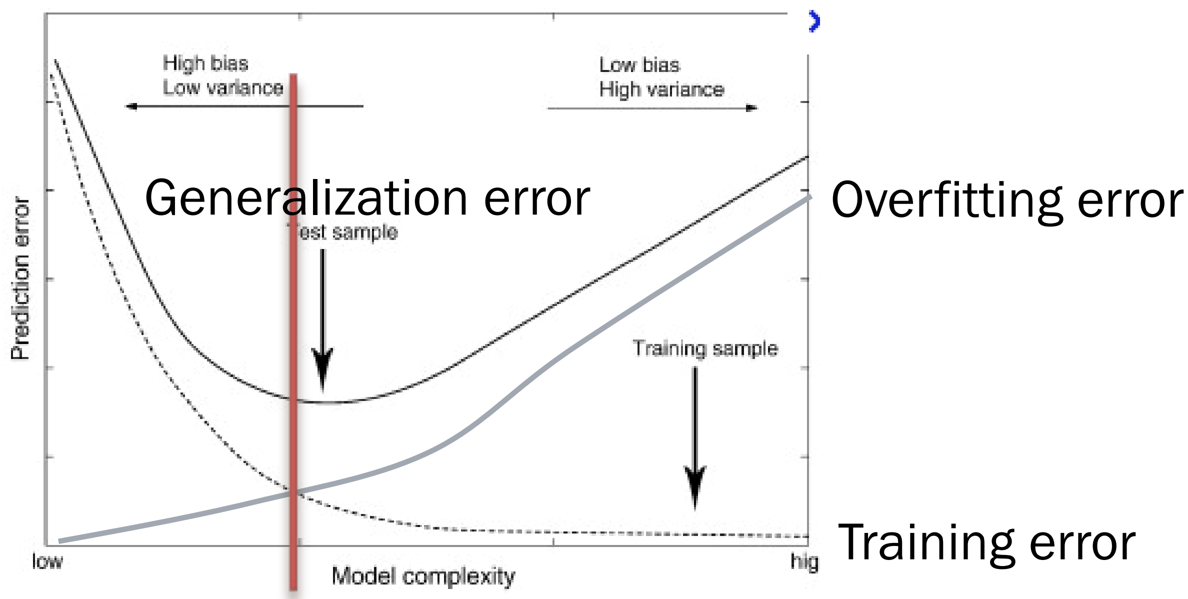


Figure 18.1 **FILES: figures/xy-plot.eps (Tue Nov 3 16:24:13 2009).** (a) Example $(x, f(x))$ pairs and a consistent, linear hypothesis. (b) A consistent, degree-7 polynomial hypothesis for the same data set. (c) A different data set, which admits an exact degree-6 polynomial fit or an approximate linear fit. (d) A simple, exact sinusoidal fit to the same data set.

OCKHAM'S RAZOR

- Given two models of similar generalization errors, one should prefer the simpler model over the more complex model
- In general there is a trade off between complex hypothesis that fit the training data well and simpler hypothesis that may generalize better.

Choosing between consistent hypothesis



EVALUATING AND CHOOSING THE BEST HYPOTHESIS

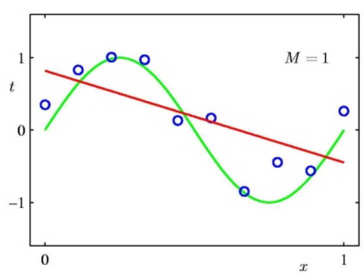
- Goal: Learn a hypothesis that fits the future data best.
  - How do we define "Future data" and "best"

- "Future data"
  - Stationary assumption: there is a prob. distribution over examples that remains stationary over time.
  - Data are selected independent and identically distributed (i.i.d)
    - P(Ej | Ej-1, Ej-2, …) = P(Ej) `independent`
    - P(Ej) = P(Ej-1) = P(Ej-2) = … `identically distributed`
  - Can use any past data as the future data for testing

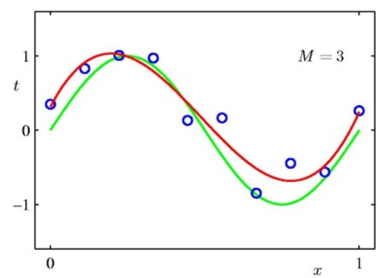
- "Best fit"
  - Error rate: proportion of mistakes it makes

MODEL SELECTION

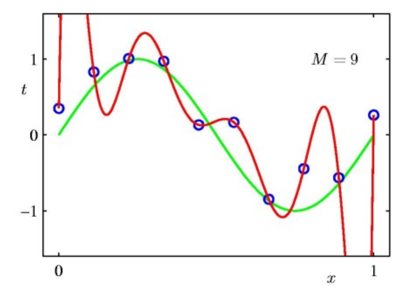
- Task of finding best hypothesis
  - Model selection: choosing hypothesis space
    - Ex> choosing the degree of the polynomial



1st Order Polynomial



3rd Order Polynomial



9th Order Polynomial

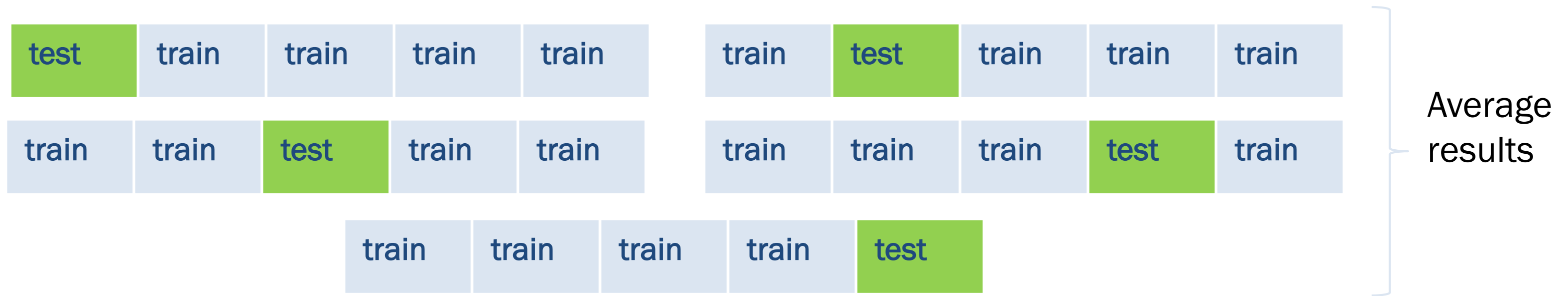
  - Optimization: finding best hypothesis within that space
    - Ex> choosing the slopes (parameter) of polynomials

- Validation set

Training set (80%)	Validation set (10%)	Test set (10%)

CROSS-VALIDATION

- k-fold cross-validation



- Leave-one-out cross-validataion (k=N)

Good tutorial: http://www.youtube.com/watch?v=hihuMBCuSIU

SIMPLE MODEL SELECTION ALGO

function CROSS-VALIDATION-WRAPPER($Learner, k, examples$) **returns** a hypothesis

  local variables: $errT$, an array, indexed by $size$, storing training-set error rates
    $errV$, an array, indexed by $size$, storing validation-set error rates
  for $size = 1$ **to** ∞ **do**
    $errT[size], errV[size] \leftarrow$ CROSS-VALIDATION($Learner, size, k, examples$)
    if $errT$ has converged **then do**
      $best\_size \leftarrow$ the value of $size$ with minimum $errV[size]$
      return $Learner(best\_size, examples)$

---

function CROSS-VALIDATION($Learner, size, k, examples$) **returns** two values:
  average training set error rate, average validation set error rate

  $fold\_errT \leftarrow 0$; $fold\_errV \leftarrow 0$
  for $fold = 1$ to k do
    $training\_set, validation\_set \leftarrow$ PARTITION($examples, fold, k$)
    $h \leftarrow Learner(size, training\_set)$
    $fold\_errT \leftarrow fold\_errT$ + ERROR-RATE($h, training\_set$)
    $fold\_errV \leftarrow fold\_errV$ +ERROR-RATE($h, validation\_set$)
  return $fold\_errT/k$, $fold\_errV/k$

FROM ERROR RATE TO LOSS

- Error rate
  - Count(y ! = $\hat{y}$)/N
- Loss function
  - L(x, y, $\hat{y}$) = Utility(result of using y given an input x) – Utility(result of using $\hat{y}$ given input x)

Generalization loss for a hypothesis h w.r.t L is

$$GenLoss_L(h) = \sum_{(x,y)\in ALLpossible\ input} L(y, h(x))P(x,y)$$

Prior prob.
Unknown

$$h^* = argmin_{h\in H} GenLoss_L(h)$$

Empirical loss

$$EmpLoss_{L,E}(h) = \frac{1}{N}\sum_{(x,y)\in E} L(y, h(x))$$

REGULARIZATION

- Doing model selection and optimization at once
  - Search for a hypothesis that directly minimized the weighted sum of empirical loss and the complexity of the hypothesis

Need to learn this para. on validation set

$$Cost(h) = EmpLoss(h) + \lambda\, Complexity(h)$$

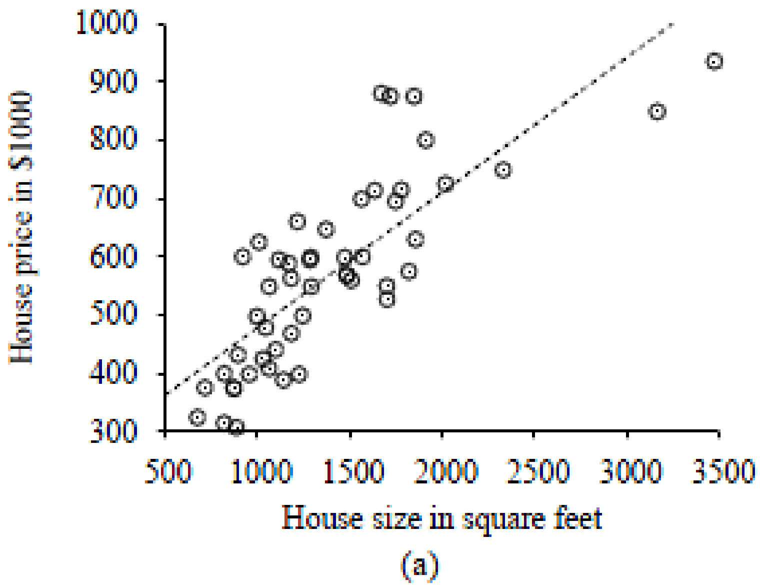
$$\hat{h}^* = argmin_{h \in H}\, Cost(h)$$

AIMA Chapter 18 (6)

PARAMETRIC LEARNING
- REGRESSION & CLASSIFICATION

UNIVARIATE LINEAR REGRESSION

- Linear regression assumes that the expected value of the output given an input, E[y|x], is linear.
- Goal of linear regression:
  - Find the gest fit h_w that minimize the loss function



$$h_w = w_1 x + w_0$$

$$Loss(h_w) = \sum_{j=1}^{N}(y_i - (w_1 x + w_0))^2$$

L2 Loss function

$$w^* = argmin_w Loss(h_w)$$

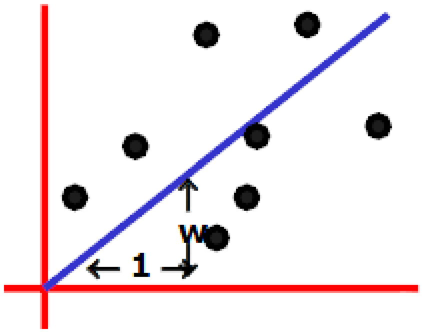
$$\Rightarrow \frac{\partial}{\partial w_0}\sum_{j=1}^{N}(y_i - (w_1 x + w_0))^2 = 0$$

$$\frac{\partial}{\partial w_1}\sum_{j=1}^{N}(y_i - (w_1 x + w_0))^2 = 0$$

UNIVARIATE LINEAR REGRESSION IN PROBABILISTIC MODEL

Assume that the data is formed by

$$y_i = wx_i + \text{noise}_i$$



inputs	outputs
$x_1 = 1$	$y_1 = 1$
$x_2 = 3$	$y_2 = 2.2$
$x_3 = 2$	$y_3 = 2$
$x_4 = 1.5$	$y_4 = 1.9$
$x_5 = 4$	$y_5 = 3.1$

where...

- the noise signals are independent
- the noise has a normal distribution with mean 0 and unknown variance σ^2

Than $P(y|w,x)$ has a normal distribution with

- mean wx
- variance σ^2

$$P(y|w, x) = \text{Normal}(\mu = wx, \sigma^2)$$

BAYESIAN LINEAR REGRESSION

$$y_i = \mathrm{w}x_i + \mathrm{Normal}(0, \sigma^2)$$

$$\mathrm{P(y|w, x)} = \mathrm{Normal}(\mu = wx, \sigma^2)$$

We have a set of datapoints (x_1,y_1) (x_2,y_2) … (x_n,y_n)
which are EVIDENCE about w.

We want to infer *w* from the data

$$\mathrm{P(w|}x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n) = \mathrm{Normal}(\mu = wx, \sigma^2)$$

- You can use BAYES rule to work out a posterior distribution for w given the data.
- Or you could do Maximum Likelihood Estimation

* BEYOND LINEAR MODEL

- Beyond linear models, analytical solutions generally do not exist:
  - require alternative method
- Maximum Likelihood Extimate
- Gradient Decent method (*note: hill climbing)

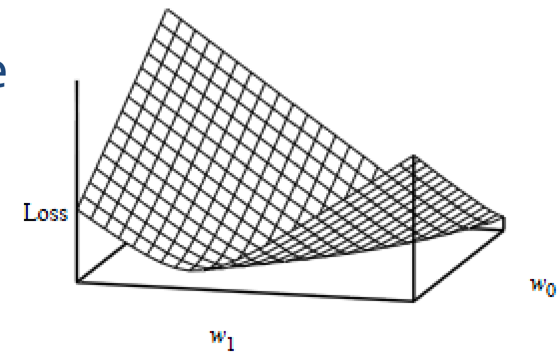
w <- any point in the parameter space

Loop until convergence do

  for each *wi* in **w** do

$$w_i \leftarrow w_i - \alpha \frac{\partial}{\partial w_i} Loss(\boldsymbol{w})$$

Step size: learning rate



TWO TYPES OF GRADIENT DECENT

- Batch gradient decent
  - Minimize the sum of the individual losses for each example

$$w_i <- w_i - \alpha \sum_j \frac{\partial}{\partial w_i}(y_i - (w_1 x_j + w_0))^2$$

  - Convergence t unique global minimal guaranteed (in linear case) as long as α is selected small enough
  - Can be slow to converge
- Stochastic gradient decent
  - Minimize the individual losses one example at a time

$$w_i <- w_i - \alpha \frac{\partial}{\partial w_i}(y_i - (w_1 x_j + w_0))^2$$

  - Fixed rate α does not guaranty convergence (but scheduled decreasing learning rate does)
  - Convergence is faster than batch gradient decent

MULTIVARIATE LINEAR REGRESSION

- Parameter optimization in multivariate regression
- AIMA Ch 18.6.2

$$h_{sw}(\mathrm{x}_j) = w_0 + w_1 x_{j,1} + w_2 x_{j,2} + \ \dots + w_n x_{j,n}$$

$$= \sum_{i=0}^{n} w_i x_{j,i} \ ; \ where \ x_{j,0} = 1$$

$$h_{sw}(\mathrm{x}_j) = \mathrm{w}^{\mathrm{T}} \mathrm{x}_j = \sum_{i=0}^{n} w_i x_{j,i}$$

$$w^* = argmin_w Loss(h_w)$$

$$Loss(h_w) = \sum_{j=1}^{N} (y_i - (w_1 x + w_0))^2$$

- Regulating the complexity in multivariate regression

$$Cost(h) = EmpLoss(h) + \lambda \, Complexity(h)$$

$$\boxed{Complexity(h_{sw}) = L_q(\mathrm{w}) = \sum_i |w_i|^q}$$

If q=1: L1 regularization -> tends to create sparse model

MULTIVARIATE LINEAR REGRESSION CONT.

Minimizing $EmpLoss(\mathrm{w}) + \lambda\, Complexity(\mathrm{w})$
= minimizing $EmpLoss(\mathrm{w})$ subjected to the constraint that
$Complexity(\mathrm{w}) \leq c$ for c that is related to λ

L1 regularization vs L2 regularization

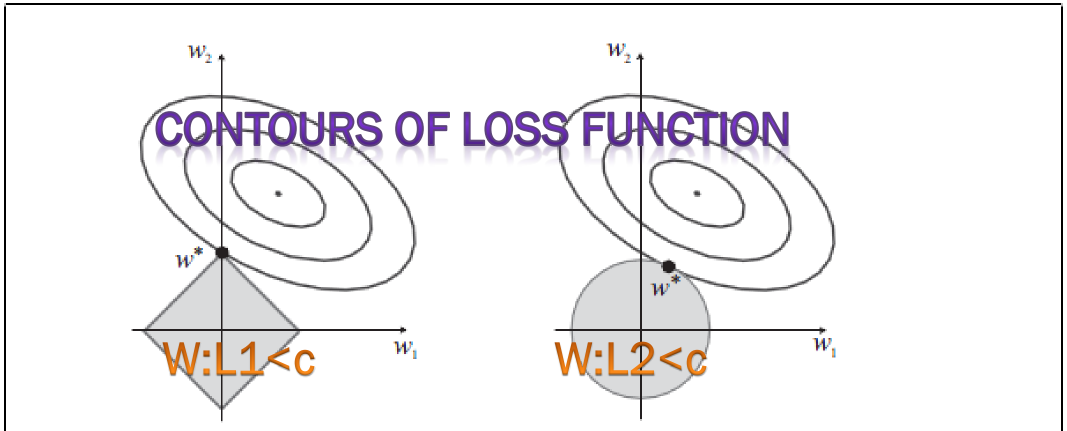


Figure 18.14 FILES: figures/diamond.eps (Wed Nov 4 14:45:53 2009). Why L_1 regularization tends to produce a sparse model. (a) With L_1 regularization (box), the minimal achievable loss (concentric contours) often occurs on an axis, meaning a weight of zero. (b) With L_2 regularization (circle), the minimal loss is likely to occur anywhere on the circle, giving no preference to zero weights.

Optimization with regularization:
Find the point closes to the minimum (center of contour)

-> L1 regularization leads to weights of zeros

LINEAR CLASSIFICATION WITH A HARD THRESHOLD

Linear functions can be used for classification as well as regression

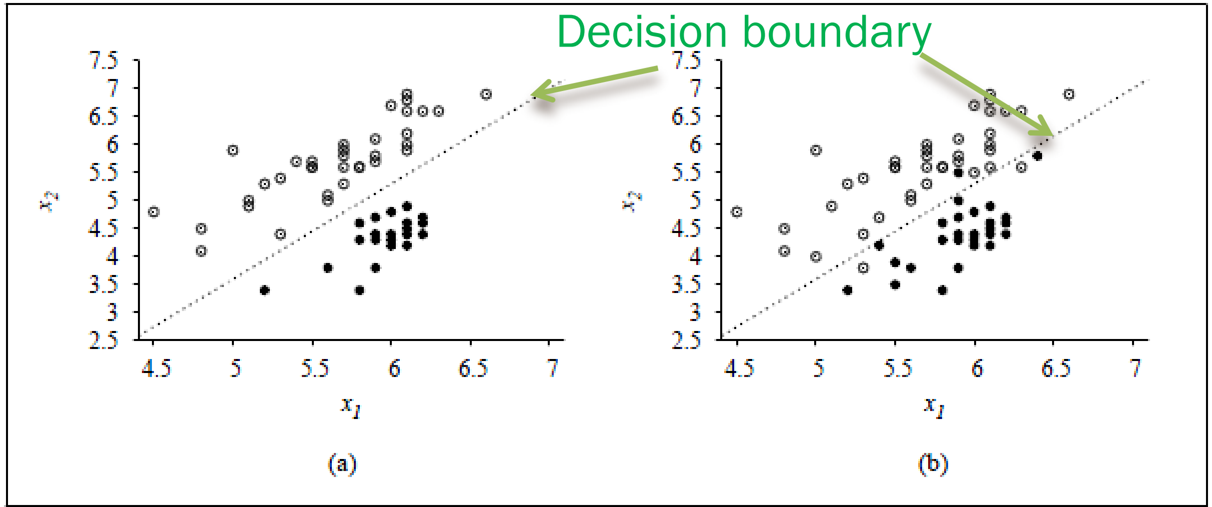


Figure 18.15 **FILES: .** (a) Plot of two seismic data parameters, body wave magnitude x_1 and surface wave magnitude x_2, for earthquakes (white circles) and nuclear explosions (black circles) occurring between 1982 and 1990 in Asia and the Middle East (?). Also shown is a decision boundary between the classes. (b) The same domain with more data points. The earthquakes and explosions are no longer linearly separable.

Decision boundary learning

LINEAR CLASSIFICATION WITH A HARD THRESHOLD CONT.

1. Classification hypothesis:

$$h_{\mathrm{w}}(\mathrm{x}) = 1 \text{ if } \mathrm{w}^T\mathrm{x} \ \geq 0$$
$$\text{otherwise } 0.$$

2. Hard threshold function:

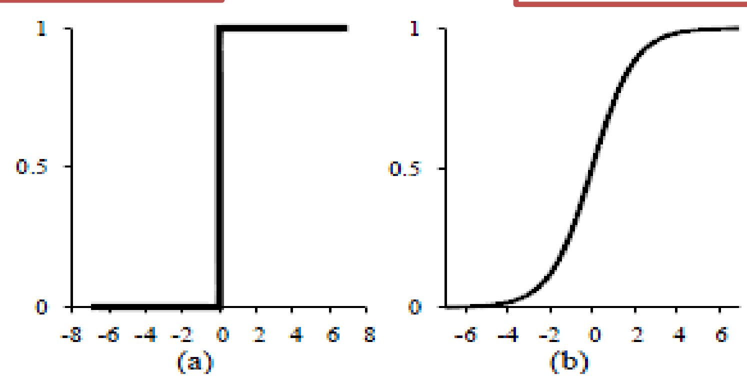
$$h_{\mathrm{w}}(\mathrm{x}) = Threshold(\mathrm{w}^T\mathrm{x}) \quad \text{where } Threshold(z) = 1 \text{ if } z \ \geq 0$$
$$\text{otherwise } 0.$$

3. Logistic:

$$h_{\mathrm{w}}(\mathrm{x}) = Logistic(\mathrm{w}^T\mathrm{x}) = \frac{1}{1+e^{-(\mathrm{w}^T\mathrm{x})}}$$

hard threshold

logistics



No longer 0/1

LOGISTIC REGRESSION: CLASSIFICATION

Unlike linear regression, it is <u>not possible to find a closed-form solution</u> for the coefficient values that maximizes the likelihood function, so an iterative process must be used. We will use **gradient decent** again.

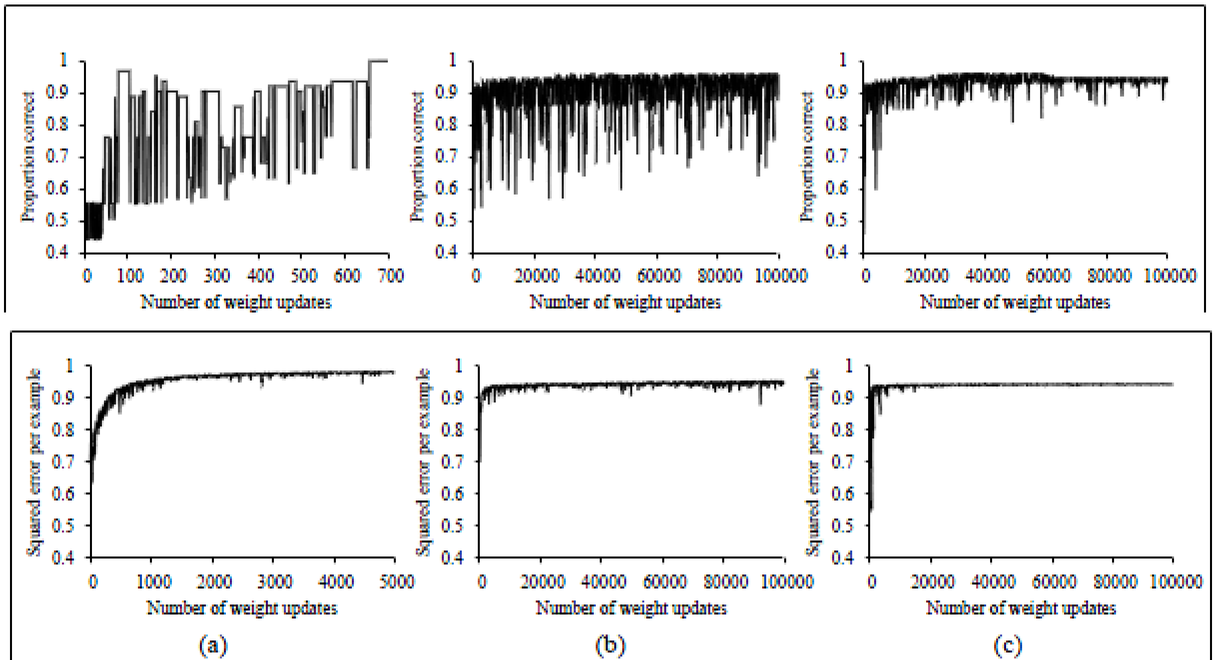


Figure 18.18 **FILES:** . Repeat of the experiments in Figure 18.15 using logistic regression and squared error. The plot in (a) covers 5000 iterations rather than 1000, while (b) and (c) use the same scale.

NAÏVE BAYES NETWORK FOR CLASSIFICATION

- Classification examples:

Spam filtering

Dear Sir.

First, I must solicit your confidence in this transaction, this is by virture of its nature as being utterly confidencial and top secret. …

TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES FOR ONLY $99

Ok, Iknow this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.

Hand written digit recognition

0

1

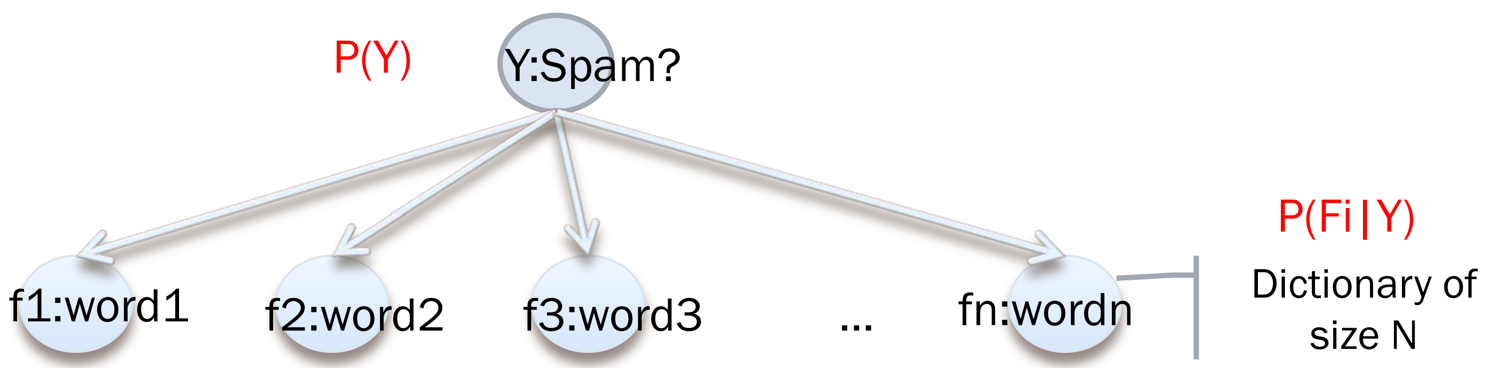
2

1

??

NAÏVE BAYES FOR SPAM FILTERING

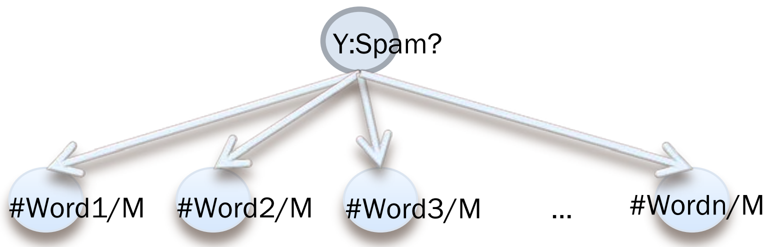
Naïve Bayes assumptions:
=> Evidences are independently drawn



P(Y,f1,f2,…,fn)= $P(Y)\prod_i P(fi|Y)$

- We only specify how each feature depends on the class
- Total number of parameters is *linear* in n

INFERENCE IN NAÏVE BAYES



MODEL

$$P(Y|f1,f2,\ldots,fn) = \alpha \frac{P(Y,f1,f2,\ldots,fn)}{P(f1,f2,\ldots,fn)}$$
$$= \alpha P(Y) \prod_i P(f_i|Y)$$

Goal: compute posterior over causes

Step 1: get joint probability of causes and evidence

$$P(Y, f_1 \ldots f_n) =$$

$$\begin{bmatrix} P(y_1, f_1 \ldots f_n) \\ P(y_2, f_1 \ldots f_n) \\ \vdots \\ P(y_k, f_1 \ldots f_n) \end{bmatrix} \Rightarrow \begin{bmatrix} P(y_1) \prod_i P(f_i|y_1) \\ P(y_2) \prod_i P(f_i|y_2) \\ \vdots \\ P(y_k) \prod_i P(f_i|y_k) \end{bmatrix}$$

Step 2: get probability of evidence $P(f1,f2,\ldots,fn)$

Step 3: renormalize $P(Y|f1,f2,\ldots,fn)$

- Assumptions:
  - Bag of words
  - Count word frequency
  - Dictionary
- What do we need in order to use naïve Bayes?
  - Estimates of local conditional probability tables
    - P(Y), the prior over labels
    - P(Fi|Y) for each feature (evidence variable)
    - These probabilities are collectively called the *parameters* of the model and denoted by θ
- Learning the **parameters**
  - P(Y) & P(Fi|Y) & P(Fi|-Y) from data

$$\sum_{i=1 \ldots size(dictionary)} P(Fi|Y) = 1 \qquad \sum_{i=1 \ldots size(dictionary)} P(Fi|-Y) = 1$$

EXAMPLE: SPAM FILTERING CONT.

SPAM

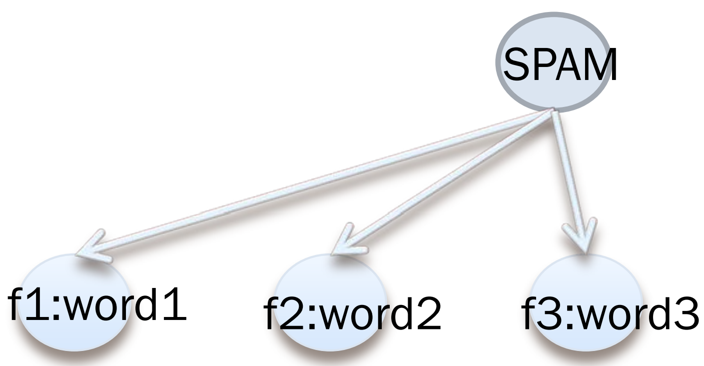
Offer is secret
Click secret link
Secret sports link

HAM

Play sports today
Went play sports
Secret sports event
Sport is today
Sport costs money

Q: P("secret" | spam) = 3/9 = 1/3
P("secret" | ham) = 1/15

P(SPAM) = 9/(9+15) = 3/8
P(HAM) = 1-3/8 = 5/8



$$P(SPAM|M=\text{"sport"})$$
$$=\frac{P(M|SPAM)P(SPAM)}{P(M|SPAM)P(SPAM)+P(M|HAM)P(HAM)}$$
$$=\frac{\frac{1}{3}*\frac{3}{8}}{\frac{1}{3}*\frac{3}{8}+\frac{1}{3}*\frac{5}{8}}$$