LASSO: Big Picture

Machine Learning – CSE446

~~Carlos Guestrin~~ Emily Fox

University of Washington

April ~~10~~ 15, 2013

©Carlos Guestrin 2005-2013

Sparsity

- Vector **w** is sparse, if many entries are zero:

  $w = (0, 0, 0, 2, 0, 0 \ldots, 1.7, 0, 0, -3, \ldots 0)$

- Very useful for many tasks, e.g.,
  - **Efficiency**: If size(**w**) = 100B, each prediction is expensive: $h(x) = w_0 + \sum w_i h_i(x)$ 100B mult.
    - If part of an online system, too slow
    - If **w** is sparse, prediction computation only depends on number of non-zeros ≪100B
  - **Interpretability**: What are the relevant dimension to make a prediction?
    - E.g., what are the parts of the brain associated with particular words?

helps understand how the brain works

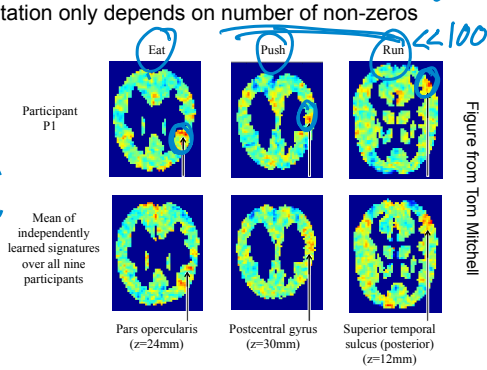


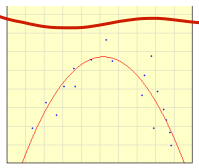
Figure from Tom Mitchell

©Carlos Guestrin 2005-2013

Regularization in Linear Regression

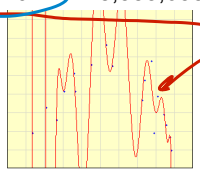
- Overfitting usually leads to very large parameter choices, e.g.:

$-2.2 + 3.1\,X - 0.30\,X^2$



$-1.1 + 4{,}700{,}910.7\,X - 8{,}585{,}638.4\,X^2 + \ldots$

penalty for large weights



overfitting

- ***Regularized*** or ***penalized*** regression aims to impose a "complexity" penalty by penalizing large weights
  - "Shrinkage" method

L_2 regularization → penalizes towards smoother functions

"ridge reg."

©Carlos Guestrin 2005-2013

LASSO Regression

$\lambda > 0$

- **LASSO:** least absolute shrinkage and selection operator
- New objective:

$$\min_{w} \sum_{j=1}^{N} \left(t(x_j) - \left(w_0 + \sum_i w_i h_i(x_j) \right) \right)^2 + \lambda \sum_{i=1}^{K} |w_i|$$

don't regularize w_0

$\|w\|_2 \longrightarrow \|w\|_1$

©Carlos Guestrin 2005-2013

Coordinate Descent for LASSO (aka Shooting Algorithm)

- Repeat until convergence
  - Pick a coordinate *j* at (random or sequentially) round robin

homework

- Set:

$$\hat{w}_\ell = \begin{cases} (c_\ell + \lambda)/a_\ell & c_\ell < -\lambda \\ 0 & c_\ell \in [-\lambda, \lambda] \\ (c_\ell - \lambda)/a_\ell & c_\ell > \lambda \end{cases}$$

- Where:

$$a_\ell = 2\sum_{j=1}^{N} (h_\ell(\mathbf{x}_j))^2$$

$$c_\ell = 2\sum_{j=1}^{N} h_\ell(\mathbf{x}_j)\left(t(\mathbf{x}_j) - (w_0 + \sum_{i \neq \ell} w_i h_i(\mathbf{x}_j))\right)$$

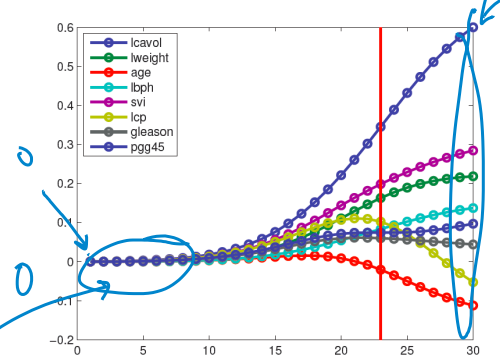
w_0??
no regularization.
$w_0 = c_0 / a_0$
$\Rightarrow w_0 = \frac{1}{N}\sum_{j=1}^{N}\left(t(x_j) - \sum_{i=1}^{K} w_i h_i(x_j)\right)$

  - For convergence rates, see Shalev-Shwartz and Tewari 2009
- Other common technique = LARS
  - Least angle regression and shrinkage, Efron et al. 2004

©Carlos Guestrin 2005-2013

Recall: *Ridge Coefficient Path*

sometimes LS = MLE



0

w_j not exactly 0

inc λ

From Kevin Murphy textbook

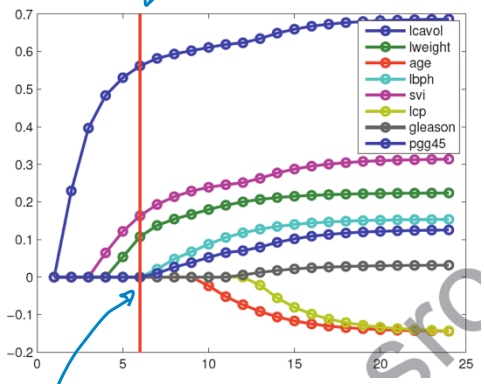
- Typical approach: select λ using cross validation

©Carlos Guestrin 2005-2013

Now: *LASSO Coefficient Path*

$\|w\|_1$

CV



3 non-zero coef.

inc λ

From Kevin Murphy textbook

LASSO Example

CV λ

Term	Least Squares	Ridge	Lasso
Intercept	2.465	2.452	2.468
lcavol	0.680	0.420	0.533
lweight	0.263	0.238	0.169
age	−0.141	−0.046	
lbph	0.210	0.162	0.002
svi	0.305	0.227	0.094
lcp	−0.288	0.000	
gleason	−0.021	0.040	
pgg45	0.267	0.133	

w_0

w_1

$\vdots$

w_k

From Rob Tibshirani slides

not in the model $\hat{w}_j = 0$

What you need to know

- Variable Selection: find a sparse solution to learning problem
- L_1 regularization is one way to do variable selection
  - Applies beyond regressions
  - Hundreds of other approaches out there
- LASSO objective non-differentiable, but convex ➔ Use subgradient
- No closed-form solution for minimization ➔ Use coordinate descent
- Shooting algorithm is very simple approach for solving LASSO

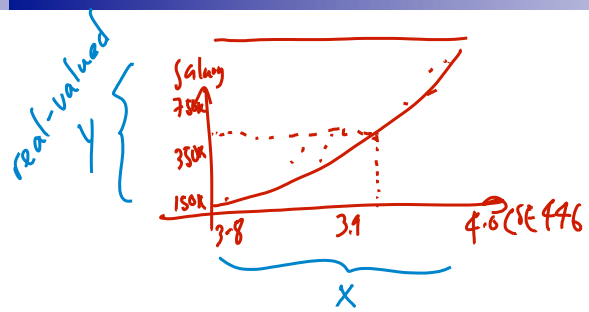
©Carlos Guestrin 2005-2013

Classification
Logistic Regression

Machine Learning – CSE446
Carlos Guestrin
University of Washington

April 15, 2013

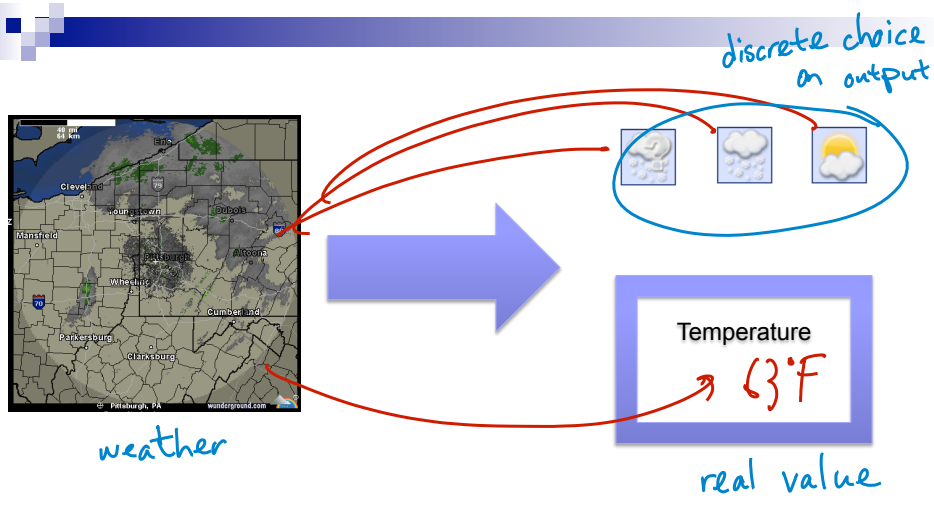
©Carlos Guestrin 2005-2013



THUS FAR, REGRESSION: PREDICT A CONTINUOUS VALUE GIVEN SOME INPUTS

©Carlos Guestrin 2005-2013

Weather prediction revisted



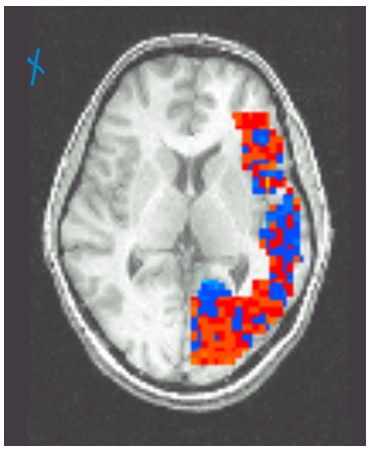
©Carlos Guestrin 2005-2013

Reading Your Brain, Simple Example

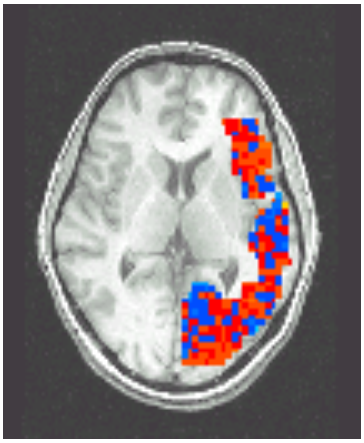
[Mitchell et al.]

Pairwise classification accuracy: 85%

Person



Animal



©Carlos Guestrin 2005-2009

Classification

$X \equiv$ (GPA, ML grade)

in reg.: Y = salary (cont. value)

now: Y is discrete

e.g. $Y \equiv$ {hired, not hired}

- **Learn**: h:**X** ↦ Y
  - **X** – features
  - Y – target classes
- Conditional probability: P(Y|**X**)

  $P(Y = \text{hired} \mid GPA = 3.6,\ ML\ grade = 3.9)$
- Suppose you know P(Y|**X**) exactly, how should you classify?
  - Bayes optimal classifier:

  $$\hat{y} = \arg\max_{y} P(Y = y \mid X = x)$$

  $P(\text{hired} \mid 3.6, 3.9) = 0.8$

  $P(\text{not hired} \mid 3.6, 3.9) = 0.2$

  $\Rightarrow$ predict hired
- **How do we estimate P(Y|X)?**

©Carlos Guestrin 2005-2013

Link Functions

- Estimating P(Y|**X**): Why not use standard linear regression?

$P(Y|X) = \underbrace{w_0 + \sum_i w_i h_i(x)}_{\text{range } (-\infty, \infty)}$?

Recall: $P(Y_j|X) \in [0,1]$

BAD

$\mathbb{R}$

- Combing regression and probability?
  - Need a mapping from real values to [0,1]
  - A link function!

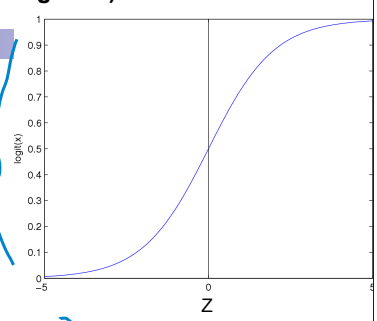
$g: \mathbb{R} \to [0,1]$

Many options, but here's a simple choice

©Carlos Guestrin 2005-2013

Logistic Regression

Logistic function (or Sigmoid): $\frac{1}{1+exp(-z)}$



$(0,1)$

$\mathbb{R}$

- Learn P(Y|**X**) directly
  - Assume a particular functional form for link function
  - Sigmoid applied to a linear function of the input features:

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

choice

z: linear just like in reg.

$w_0 + \sum_i w_i x_i$ ← not bounded, could be neg.

after logistic fcn, output is in $[0,1]$

Features can be discrete or continuous!

©Carlos Guestrin 2005-2013

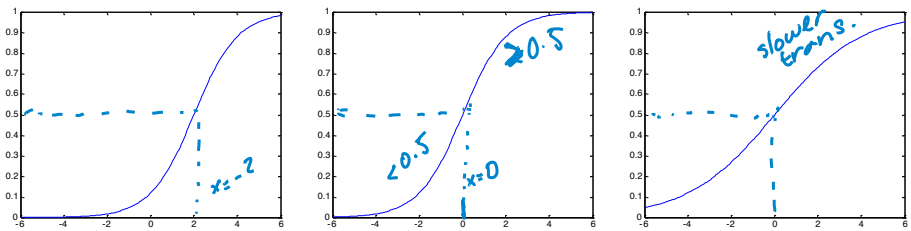
Understanding the sigmoid

$$g(w_0 + \sum_i w_i x_i) = \frac{1}{1 + e^{w_0 + \sum_i w_i x_i}}$$

$k=1$

shift

w_0=-2, w_1=-1 | w_0=0, w_1=-1 | w_0=0, w_1=-0.5

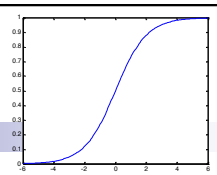


$z = w_0 + w_1 x$

©Carlos Guestrin 2005-2013

Logistic Regression – a Linear classifier

$$\frac{1}{1 + exp(-z)}$$



$P(Y=0|x,w) =$

$$g(w_0 + \sum_i w_i x_i) = \frac{1}{1 + e^{w_0 + \sum_i w_i x_i}}$$

$w_0 + \sum_i w_i x_i > 0$

$\Rightarrow g(w_0 + \sum_i w_i x_i) < 0.5$

$\Rightarrow P(Y=0|x,w) < 0.5$

$\Rightarrow$ predict class = 1

$w_0 + \sum_i w_i x_i = 0$

$w_0 + \sum_i w_i x_i < 0$

$\Rightarrow g(w_0 + \sum_i w_i x_i) > 0.5$

$\Rightarrow$ predict class = 0

©Carlos Guestrin 2005-2013

Very convenient!

2 - class $P(Y=1|x) = 1 - P(Y=0|x)$

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

$\ln 1 = 0$

$$P(Y=0\,|X=<X_1,...X_n>) = \frac{1}{1+exp(w_0+\sum_i w_i X_i)}$$

implies

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

implies ratio of prob.

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

$\Rightarrow$ class 1 wins else class 0

linear classification rule!

implies

log odds $$\ln\frac{P(Y=1\,|X)}{P(Y=0\,|X)} = \underbrace{w_0+\sum_i w_i X_i}_{\text{linear}} > 0$$

$\Rightarrow$ class 1 wins as before

©Carlos Guestrin 2005-2013

Loss function: Conditional Likelihood

$\ln \prod = \sum \ln$ ln is monotone

- Have a bunch of iid data of the form: $(x^j, y^j)_{1:N} = D = (D_x, D_y)$ iid

  x y

  (GPA: 3.2 hired)

  (3.4, 3.8, not hired)

- Discriminative (logistic regression) loss function:
  Conditional Data Likelihood

$$\arg\max_w P(D_y \mid D_x, w) = \arg\max_x \prod_{j=1}^N P(y^j \mid x^j, w)$$

$$= \arg\max_w \ln \prod_{j=1}^N P(y^j \mid x^j, w) = \arg\max_x \sum_{j=1}^N \ln P(y^j \mid x^j, w)$$

$$\ln P(\mathcal{D}_Y \mid \mathcal{D}_{\mathbf{X}}, \mathbf{w}) = \sum_{j=1}^N \ln P(y^j \mid \mathbf{x}^j, \mathbf{w})$$

©Carlos Guestrin 2005-2013

Expressing Conditional Log Likelihood

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

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

$$l(\mathbf{w}) \equiv \sum_j \ln P(y^j|\mathbf{x}^j,\mathbf{w})$$

$$= \sum_j \begin{cases} P(Y=1|x_j,w) & \text{if } y^j=1 \\ P(Y=0|x^j,w) & \text{if } y^j=0 \end{cases}$$

$$\ell(\mathbf{w}) = \sum_j y^j \ln P(Y=1|\mathbf{x}^j,\mathbf{w}) + (1-y^j)\ln P(Y=0|\mathbf{x}^j,\mathbf{w})$$

$$= \sum_j y^j \ln \frac{\exp(w_0+\sum_i w_i x_i)}{1+\exp(w_0+\sum_i w_i x_i)} + (1-y^j)\ln\frac{1}{1+\exp(w_0+\sum_i w_i x_i)}$$

$$= \sum_j y^j (w_0+\sum_i w_i x_i) - \ln\left(1+\exp(w_0+\sum_i w_i x_i)\right)$$

©Carlos Guestrin 2005-2013

Maximizing Conditional Log Likelihood

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

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

$$l(\mathbf{w}) \equiv \ln \prod_j P(y^j|\mathbf{x}^j,\mathbf{w})$$

$$= \sum_j y^j(w_0+\sum_i^n w_i x_i^j) - \ln(1+exp(w_0+\sum_i^n w_i x_i^j))$$

Good news: *l*(**w**) is concave function of **w**, no local optima problems

Bad news: no closed-form solution to maximize *l*(**w**)

Good news: concave functions easy to optimize

©Carlos Guestrin 2005-2013