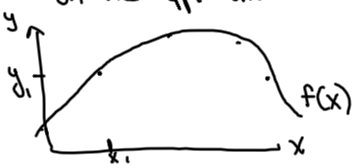
Dr. Cynthia Rudin (CCLS, Columbia)

Regression (the Statistical Learning Version anyway):

Given $\{(x_i, y_i)\}_{i=...m}$ where $x_i \in X$ & $y_i \in \mathbb{R}$, chosen randomly from an unknown probability distribution, find $f \in F$ $f: X \to \mathbb{R}$, such that for a new randomly chosen instance $x \in X$, we have $f(x)$ close to its label y.

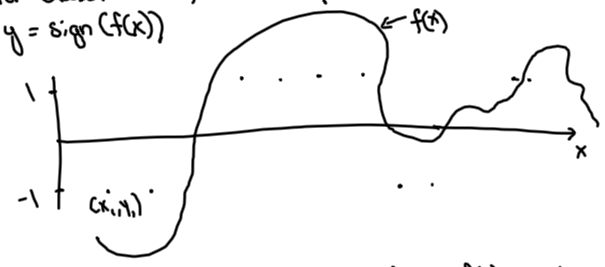
in some sense which varies on the application

Curve fitting



Compare Classification & Regression

For classification, want to produce f(x) such that $y = \text{sign}(f(x))$



For regression, want to produce f(x) such that $y = f(x)$

Examples:

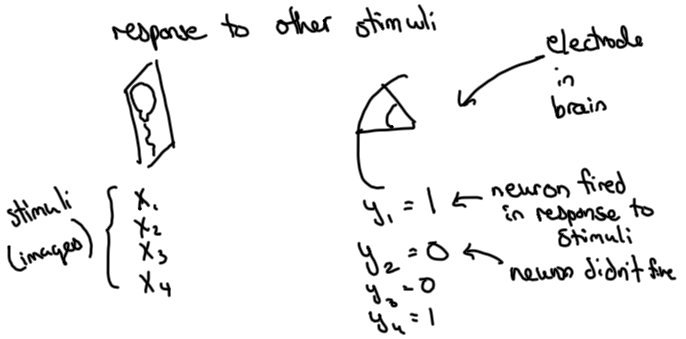
1) Want to estimate the number of customers for a given product, given data about the product and data about past product sales.

2) Time Series Analysis, for instance, predicting the price of stocks, or predicting the path of a moving object

3) Cancer Genomics (Hans & West 06)
Use gene expression levels (microarray data) from tumor samples to estimate:
- survival time
- amount of lymph node invasion
- response to a particular treatment

Jan 31-2:36 PM

4) Computational Neuroscience: Calculating the firing rate of a neuron in the visual cortex for a new stimuli, based on the neuron's response to other stimuli



firing rate: prob that neuron will fire in response to stimulus x.

$= \mathbb{P}(y=1 \mid x) = \mathbb{E}(y|x) = \int y \, du(x,y)$

(# of dimensions = # of pixels of images)

Square Loss and the Mean:

We generally would like to choose $f(x)$ so that it is close to y in some sense, for instance, we might want to choose f to minimize:

$\mathbb{E}[(y - f(x))^2 \mid x]$ for each x

Say $y \in \{0, 1\}$. Fix x. What is the best $h = f(x)$? Let $p = \mathbb{P}(y=1|x)$. Turns out the answer is $h = p$, the mean.

We cannot minimize $\mathbb{E}[(y-f(x))^2]$ in practice, because we only have a finite sample of the data $\{(x_i, y_i)\}_{i=1..m}$

In practice, we minimize an "empirical" (based on data) quantity:

$\frac{1}{m} \sum_{i=1}^{m} (y_i - f(x_i))^2$

http://www1.cs.columbia.edu/~allen/S08/NOTES/linkedlists.pdf

Legendre published & named least squares in 1805 but Gauss claimed he had known the method since 1795. They both used it to determine from astronomical observation the orbits of bodies around the sun.

Solving Least Sq & Pseudo Inverse

Choose f to be linear. $f(\vec{x}) = \vec{w} \cdot \vec{x}$ for $x \in \mathbb{R}^n$

$= \sum_j w_j x_j$

Least Sq Problem: $\min_{w \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^{m} (\vec{w} \cdot \vec{x}_i - y_i)^2$

Jan 31-2:56 PM

Least ∂_8: $\min_{\vec{\omega} \in \mathbb{R}^n} \underbrace{\frac{1}{m} \sum_{i=1}^{m} (\vec{\omega} \cdot \vec{x}_i - y_i)^2}_{\Phi(\vec{\omega})}$

To solve, set $\nabla \Phi(\omega) = 0$

$$\nabla \Phi(\vec{\omega}) = \begin{pmatrix} d\Phi/d\omega_1 \\ \vdots \\ d\Phi/d\omega_n \end{pmatrix} = \frac{1}{m} 2 \underline{\underline{X}}^T (\underline{\underline{X}} \cdot \vec{\omega} - \vec{y})$$

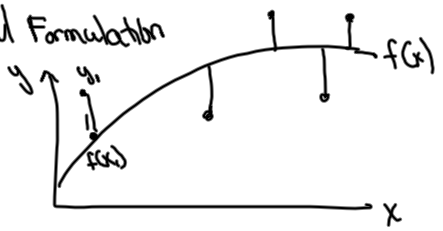
where $\underline{\underline{X}} = \begin{pmatrix} — & x_1^T & — \\ & \vdots & \\ — & x_m^T & — \end{pmatrix}$

$$\nabla \phi(\vec{\omega}) = 0 \Rightarrow \vec{\omega} = \underbrace{(\underline{\underline{X}}^T \cdot \underline{\underline{X}})^{-1} \underline{\underline{X}}^T}_{\text{pseudoinverse}} y$$

If colums of $\underline{\underline{X}}$ are indep., pseudoinverse exists.
If $\underline{\underline{X}}$ has lin. indep columns & rows, $\underline{\underline{X}}^{-1}$ is the pseudoinverse

Pseudoinverse: Fredholm 1903
Moore 1920, Penrose 1955

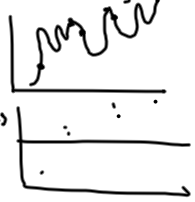
General Formulation



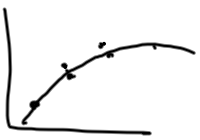
For regression (& other learning problems) we often try to minimize a function of the form

$$L(f) = \underbrace{\sum_{i=1}^{m} l(y_i, f(x_i))}_{\text{"empirical error" term measures closeness of data to model}} + \underbrace{c}_{\substack{\text{regularization} \\ \text{parameter} \\ \in \mathbb{R}}} \underbrace{R(f)}_{\substack{\text{regularization term measures smoothness} \\ \text{of } f \text{ to prevent} \\ \text{overfitting}}}$$

If c is too small, model may overfit
If c is too large, " " underfit



If c is just right, model generalizes (predicts) well



Some loss functions for regression:

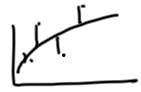
Jan 31-3:18 PM

Jan 31-3:41 PM

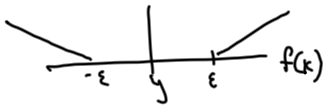
Jan 31-3:41 PM

least squares (l_2): $l(y, f(x)) := (y - f(x))^2$

absolute distance (l_1): $l(y, f(x)) := |y - f(x)|$



SVM ε-insensitive $l(y, f(x)) := \begin{cases} 0 & \text{if } |y - f(x)| < \varepsilon \\ |y - f(x)| & \text{otherwise} \end{cases}$



Some regularization terms, assuming $f_w(\vec{x}) = \vec{w} \cdot \vec{x}$:

$R(f_w) = \|w\|_2^2 = \sum_j w_j^2$

$R(f_w) = \|w\|_1 = \sum_j |w_j|$

$R(f_w) = \|f\|_{RKHS}^2$ (you'll hear plenty about this later!)

Tikhonov Regularization Problem or Ridge Regression:

$\min_{\vec{w}} \|X w - y\|_2^2 + \|\Gamma w\|_2^2$ can be solved:

(Γ: pos. semidef, $z \cdot \Gamma \cdot z \geq 0 \ \forall z \in \mathbb{R}$)

$$w = (X^T X + \Gamma^T \Gamma)^{-1} X^T y$$

True l_0 sparsity

There has been a recent flurry of work (inspired by a result of Donoho) showing that in some cases (when columns of X are almost orthogonal) the minimizer of:

(*1) $\|X w - y\|^2 + C \|w\|_1$

is the same as the minimizer of the real sparsity problem:

(*2) $\|X w - y\|^2 + C \|w\|_0$

where $\|w\|_0$ = the number of nonzero elements of w.

However (*1) is convex & we can solve it whereas (*2) is not!

Linear vs. Nonlinear Modelling

There are many options for the form of the function f.

$f(\vec{x}) = \vec{w} \cdot \vec{x}$ linear

$f(\vec{x}) = \vec{w} \cdot \phi(\vec{x})$ "non-linear"

ϕ can be a vector in an infinite-dimensional Hilbert space.

This means w is infinite dimensional too!

Jan 31-2:37 PM

This intro to regression is only the beginning!

Many other techniques:

- splines
- regression trees
- RBF networks & other non linear methods

In 1900 3 dims was a lot.

In 1980's 10 dims was a lot.

Now, we have ~10,00 dim