Lecture 1. Introduction and Linear Models for Regression

Bao Wang
Department of Mathematics
Scientific Computing and Imaging Institute
The University of Utah
Math 5750/6880, Fall 2023

Topics of this course

1. Linear models for regression and classification.
2. First-order optimization methods for data science. (GD, Proximal GD, SGD, Acceleration)
3. Neural networks (CNN, RNN, GNN, Transformers).
4. Clustering and dimension reduction (k-means, spectral clustering, PCA, CS, ...).
5. Basics of learning theory. (PAC, NTK, ...)

Basic Information

Instructor: Bao Wang

Meeting Time: TuTh 12:25pm - 1:45pm, FASB 250

Office Hours: TuTh 3:00pm - 4:30pm.

Office: WEB 4652.

Email: `bwang@math.utah.edu`

Course webpage:
Math 5750: `https://utah.instructure.com/courses/890139`
Math 6880: `https://utah.instructure.com/courses/891566`

References:
1. Zhang et al., Dive into Deep Learning
2. Shai Shalev-Shwartz and Shai Ben-David, Understanding Machine Learning: From Theory to Algorithms
3. Christopher Bishop, Pattern Recognition and Machine Intelligence
4. Some recent literature

Grades

Lecture notes scribe 40%: We will present the lectures using slides and each team needs to scribe each lecture and submit one copy of the lecture notes in PDF along with the latex source. We will provide a latex template. Please submit at least scribes of five lectures.

Course projects 60%: There will be three projects in total and the lowest score will be dropped. For each project, you will need to write a report and submit it along with the source code.

Course grade:

P	90	85	80	77	70	68	66	50	45	40
Grade	A	A-	B+	B	B-	C+	C	C-	D+	D

Team: 3 students in a team.

Important Dates

Last day to **register** is Sep 1

Last day to **drop** class is Oct 20

Holidays: There will be no class on Sep 4 (Labor Day), Oct 8-15 (Fall break), Nov 23-26 (Thanksgiving).

Regression problem

Given $\{\boldsymbol{x}_n, t_n\}_{n=1}^{N}$ where $\boldsymbol{x}_n$ is the observation and t_n is the corresponding target.

How to find t that corresponds to the other $\boldsymbol{x}$?

Regression problem

Given $\{\boldsymbol{x}_n, t_n\}_{n=1}^{N}$ where $\boldsymbol{x}_n$ is the observation and t_n is the corresponding target.

How to find t that corresponds to the other $\boldsymbol{x}$?

Construct a function $y(\boldsymbol{x})$ based on the given dataset $\{\boldsymbol{x}_n, t_n\}_{n=1}^{N}$.

Linear regression model

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + w_1 x_1 + \cdots + w_D x_D, \quad \boldsymbol{x} = (x_1, \cdots, x_D)^\top. \tag{1}$$

We can extend the model (1) by considering linear combinations of fixed nonlinear functions of the input variables, of the form

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\boldsymbol{x}), \tag{2}$$

where $\phi_j(\boldsymbol{x})$ are known as basis functions.

It is often convenient to define an additional dummy 'basis function' $\phi_0(\boldsymbol{x}) = 1$ so that

$$y(\boldsymbol{x}, \boldsymbol{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\boldsymbol{x}) = \boldsymbol{w}^\top \phi(\boldsymbol{x}), \tag{3}$$

where $\boldsymbol{w} = (w_0, \cdots, w_{M-1})^\top$ and $\phi = (\phi_0, \cdots, \phi_{M-1})^\top$.

Note $y(\boldsymbol{x}, \boldsymbol{w})$ is a nonlinear function of the input vector $\boldsymbol{x}$.

Basis functions

1. *Polynomial basis:* $\phi_j(x) = x^j$.
2. *Gaussian basis:* $\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\}$, where μ_j govern the locations of the basis functions in input space, and the parameter s governs their spatial scale.
3. *Sigmoidal basis:* $\phi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right)$, where $\sigma(a) = \frac{1}{1+\exp(-a)}$ is the logistic sigmoid function.

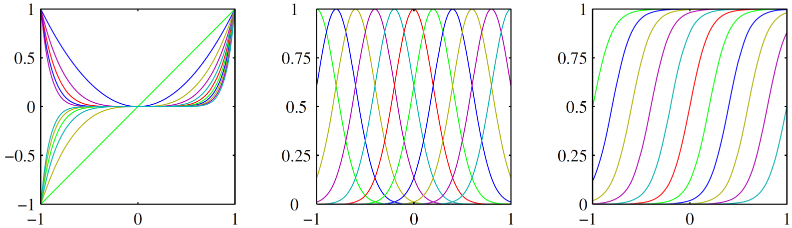

Figure: Left: polynomial basis; Middle: Gaussian basis; Right: Sigmoidal basis.

Linear regression

Given the N observation $\{\boldsymbol{x}_n, t_n\}_{n=1}^{N}$, we can define the following loss function

$$L(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(t_n - \boldsymbol{w}^\top \phi(\boldsymbol{x}_n)\right)^2, \tag{4}$$

which can be written in the following compact form

$$L(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{t} - \Phi \boldsymbol{w}\|^2, \tag{5}$$

where

$$\Phi = \begin{pmatrix} \phi_0(\boldsymbol{x}_1) & \phi_1(\boldsymbol{x}_1) & \cdots & \phi_{M-1}(\boldsymbol{x}_1) \\ \phi_0(\boldsymbol{x}_2) & \phi_1(\boldsymbol{x}_2) & \cdots & \phi_{M-1}(\boldsymbol{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\boldsymbol{x}_N) & \phi_1(\boldsymbol{x}_N) & \cdots & \phi_{M-1}(\boldsymbol{x}_N) \end{pmatrix}. \quad \text{and} \quad \boldsymbol{t} = \begin{pmatrix} t_1 \\ \vdots \\ t_N \end{pmatrix}, \tag{6}$$

where Φ is called the *design matrix*, whose elements are given by $\Phi_{nj} = \phi_j(\boldsymbol{x}_n)$.

Note $L(\boldsymbol{w})$ is a quadratic function of $\boldsymbol{w}$, let $dL(\boldsymbol{w})/d\boldsymbol{w} = 0$, we have

$$0 = \frac{dL(\boldsymbol{w})}{d\boldsymbol{w}} = \Phi^\top(\Phi\boldsymbol{w} - \boldsymbol{t}),$$

therefore,

$$\boldsymbol{w}_{ML} = (\Phi^\top\Phi)^{-1}\Phi^\top\boldsymbol{t}. \tag{7}$$

The quantity $\Phi^\dagger \equiv (\Phi^\top\Phi)^{-1}\Phi^\top$ is known as the *Moore-Penrose pseudo-inverse* of the matrix Φ.

Overfitting

How to select M?

Let us consider a given set of data, which is sampled from $\sin(2\pi x)$ with noise.

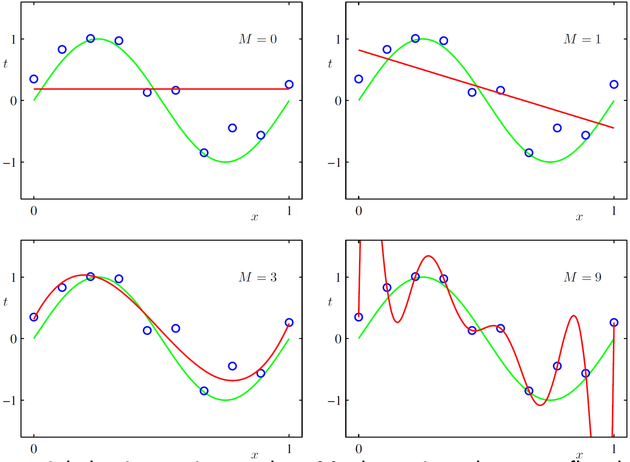

Figure: Plots of polynomials having various orders M, shown in red curves, fitted to a dataset sampled, with noise, from $\sin(2\pi x)$. Overfitting happens when M is large!

Why overfitting?

	$M = 0$	$M = 1$	$M = 6$	$M = 9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

Figure: Table of the coefficients $\boldsymbol{w}^*$ for polynomials of various order. The magnitude of the coefficients increases dramatically as the order of the polynomial increases.

Why overfitting?

	$M=0$	$M=1$	$M=6$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

Figure: Table of the coefficients $\boldsymbol{w}^*$ for polynomials of various order. The magnitude of the coefficients increases dramatically as the order of the polynomial increases.

Regularization is all you need!

L_2-regularization:

$$L^2(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\Big(t_n - \boldsymbol{w}^\top\phi(\boldsymbol{x}_n)\Big)^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2, \tag{8}$$

where $\|\boldsymbol{w}\|_2^2 \equiv \boldsymbol{w}^\top\boldsymbol{w} = w_0^2 + w_1^2 + \cdots + w_{M-1}^2$.

> λ governs the relative importance of the regularization term compared with the sum-of-squares error term. What will happen if λ is very large?

> *This is also known in the statistics literature as parameter shrinkage method because they reduce the value of the coefficients. The particular case of quadratic regularizer is called ridge regression. In the context of neural networks, this approach is known as weight decay.*

L_q-regularization:

$$\frac{1}{2}\sum_{n=1}^{N}\Big(t_n - \boldsymbol{w}^\top \phi(\boldsymbol{x}_n)\Big)^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q. \tag{9}$$

$q = 1$ corresponds to *Lasso regression.*

Lagrange Multipliers

$$\min_{\boldsymbol{x}} f(\boldsymbol{x})$$

s.t.

$$g(\boldsymbol{x}) \leq C$$

is equivalent to

$$\min_{\boldsymbol{x},\lambda} f(\boldsymbol{x}) + \lambda g(\boldsymbol{x})$$

s.t.

$$\lambda \geq 0;\ \ g(\boldsymbol{x}) \leq C;\ \ \lambda(g(\boldsymbol{x}) - C) = 0.$$

Lasso: $q = 1$ in (9), when λ is sufficiently large, some of the coefficients w_j are driven to zero, leading to a *sparse* model in which the corresponding basis functions play no role. Minimizing (9) is equivalent to minimizing the unregularized sum-of-squares error subject to the constraint

$$\sum_{j=1}^{M} |w_j|^q \leq \eta \tag{10}$$

for an appropriate value of the parameter η, where the two approaches can be related using Lagrange multipliers.

Sparsity due to the L_1-regularization

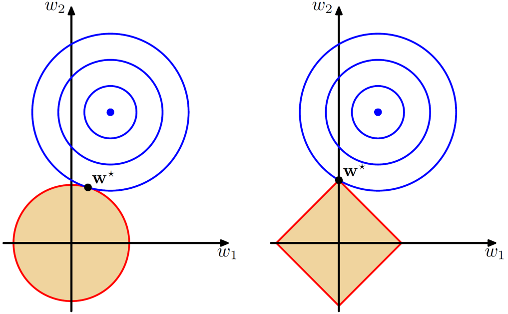

Figure: Plot of the contours of the unregularized error function (blue) along with the constraint region for the quadratic regularizer $q = 2$ on the left and the lasso regularizer $q = 1$ on the right, in which the optimum value for the parameter vector $\boldsymbol{w}$ is denoted by $\boldsymbol{w}^*$. Lasso gives $w_1^* = 0$.

Solution to the L_1-regularized regression

Lasso regularization problem has the following compact form

$$L(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{t} - \Phi\boldsymbol{w}\|_2^2 + \gamma\|\boldsymbol{w}\|_1, \tag{11}$$

where $\gamma = \lambda/2$. Consider a simple case first, where Φ is orthogonal.

Expanding (11) and excluding irrelevant terms, we have the equivalent form

$$\min_{\boldsymbol{w}} \left(-\boldsymbol{t}^\top\Phi\boldsymbol{w} + \frac{1}{2}\|\boldsymbol{w}\|^2 \right) + \gamma\|\boldsymbol{w}\|_1.$$

Solution to the L_1-regularized regression

Let $\boldsymbol{t}^\top \Phi := \beta = (\beta_0, \cdots, \beta_{M-1})$, the previous problem can be rewritten as

$$\min_{\boldsymbol{w}} \sum_{i=0}^{M-1} -\beta_i w_i + \frac{1}{2} w_i^2 + \gamma |w_i|.$$

Fix a certain i, we want to minimize

$$\mathcal{L}_i = -\beta_i w_i + \frac{1}{2} w_i^2 + \gamma |w_i|.$$

If $\beta_i > 0$, then we must have $w_i \geq 0$. why?

Solution to the L_1-regularized regression

If $\beta_i > 0$, then we must have $w_i \geq 0$. Otherwise, let $w_i^* < 0$ minimizes $\mathcal{L}_i$, then $-w_i^*$ enables even smaller $\mathcal{L}_i$.

If $\beta_i < 0$, then we must choose $w_i \leq 0$.

Solution to the L_1-regularized regression

$$\mathcal{L}_i = -\beta_i w_i + \frac{1}{2} w_i + \gamma |w_i|.$$

> If $\beta_i > 0$, since $w_i \geq 0$,

$$\mathcal{L}_i = -\beta_i w_i + \frac{1}{2} w_i^2 + \gamma w_i,$$

and differentiating this with respect to w_i and setting equal to zero, we get

$$w_i^* = \beta_i - \gamma,$$

and this is only feasible if the right-hand side is nonnegative (we require $w_i \geq 0$), so in this case the actual solution is

$$w_i^* = sgn(\beta_i)(|\beta_i| - \gamma)^+.$$ Note that $\beta_i, \gamma > 0$ and $\beta_i - \gamma \geq 0$.

Solution to the L_1-regularized regression

$$\mathcal{L}_i = -\beta_i w_i + \frac{1}{2} w_i + \gamma |w_i|.$$

> If $\beta_i \leq 0$. This implies we must have $w_i \leq 0$ and so

$$\mathcal{L}_i = -\beta_i w_i + \frac{1}{2} w_i^2 - \gamma w_i.$$

Differentiating with respect to w_i and setting equal to zero, we get

$$w_i^* = sgn(\beta_i)(|\beta_i| - \gamma).$$

But again, to ensure this is feasible, we need $w_i \leq 0$, which is achieved by taking

$$w_i^* = sgn(\beta_i)(|\beta_i| - \gamma)^+.$$

Solution to the L_1-regularized regression

$$L(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{t} - \Phi\boldsymbol{w}\|_2^2 + \gamma\|\boldsymbol{w}\|_1,$$

Shrink:

$$w_i^* = sgn(\beta_i)(|\beta_i| - \gamma)^+.$$