Linear Models

ID2214 Programming for Data Science

https://gits-15.sys.kth.se/amiakh/ID2214

Amir Hossein A. Rahnama
arahnama@kth.se

Overview: statistical learning theory

- Let X be the vector space of all inputs and Y be the vector space of all outputs.
- Statistical learning theory takes the perspective that **there is an unknown probability function** f over the product space $Z = X \times Y$ that the training set are samples from this unknown probability function.
- The learning then can be formulated as an inference of this function $f : X \rightarrow Y$ such that $f(x) = y$.
- Let $\mathcal{H}$ be the space of all those functions and $l(f(x), y)$ be the loss function that represents the difference between the predicted value of $f(x)$ and y.
- Because the probability function is unknown, we **minimize** *empirical risk*:

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{N} l(f(x_i), y_i) \tag{1}$$

Overview: Linear functions

- A linear function is a polynomial function that x has the degree of at most 1:

$$f(x) = W \cdot X + b \tag{2}$$

- The name linear refers to the fact that the set of $(x, f(x))$ in the Cartesian plane is a line.
- The slope of the function, how steeply the line is slanted, is given by W.
- In calculus, the derivative of a function measures the rate of changes in a function. In linear function, $f'(x) = W$, therefore the change of the function does not depend on the input X.

Why should we study linear models?

- They are simple, hence they provide an interpretable description of how the inputs affect the output
- When it comes to prediction, they can outperform nonlinear models, especially in the following situations:
  - Small number of training data
  - Low signal-to-noise
  - Sparse data
- They can be used to model non-linear relationships as well!

Linear Regression and Least Squares

- An input vector of the form $x = (x_1, ..., x_m) \in \mathcal{R}^m$
- The task is to predict the outcome $y \in \mathcal{R}$
- The model of choice is in form of $y = w_0 + \sum_{j=1}^{m} x_j w_j + \epsilon$
- In probabilistic language, we can write above as $p(y|x, w) = \mathcal{N}(y|\mu(x), \sigma^2(x))$.
- w_j are the *unknown* parameters or coefficients (also called regressor) that affect the outcome variable namely y (also called regressand or dependent variable)
- ϵ is the unobserved error term that has the mean of zero: $\epsilon \sim N(0, \sigma^2)$
- In the calculus language, w_j represents the change in the dependent variable when the regression changes by one unit, i.e. $\frac{\partial y}{\partial x_1} = w_1$
- In a linear model, we assume that $E(Y|X)$ is linear or is approximately linear

What inputs can we use in linear regression?

- Quantitative inputs
- Transformations of quantitative inputs, log or square of input
- Basic expansions, e.g. $x_2 = x_1^2$ that leads to polynomial representations
- Numeric, dummy or one-hot representations of qualitative or categorical inputs that are of type levels or factors
- Interaction between variables, e.g. $x_3 = x_1 \cdot x_2$

Least squares: computation

We would like to minimize the RSS (residual sum of squares) loss function:

$$\min_{w} \quad (\mathbf{Y} - \mathbf{XW})^T(\mathbf{Y} - \mathbf{XW}) \tag{3}$$

$$\frac{\partial \text{RSS}}{\partial w} = -2\mathbf{X}^T(\mathbf{y} - X\mathbf{W}) \tag{4}$$

Using Fermat's Theorem, we can find the solution by:

$$\mathbf{X}^T(\mathbf{Y} - \mathbf{XW}) = 0 \tag{5}$$

and hence

$$\hat{W} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{6}$$

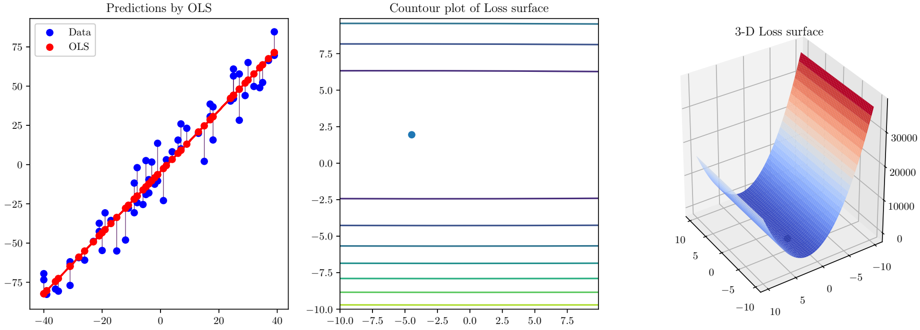
Least squares II

- Training data $(x_i, y_i) \quad i = 1, ..., N$ are independent random draws from the population
- If your training observations are **not drawn randomly**, RSS is valid if y_is are conditionally independent given the inputs x_i
- RSS has no assumption about the validity of the model it finds
- In linear model, we assume that $E(Y|X)$ is linear or is approximately linear
- Geometric interpretation: the chosen $\hat{W}$ is **orthogonal** to the residual error

Least squares: 2-D example

- Let $X \in \mathcal{R}$ be a sample of 100 numbers in $[-40, 40]$, 60 of which are used for training.
- Let us assume that $w_0 = 1$ and $w_1 = -3.5$.
- We formulated the problem as $Y = w_0 + w_1 x_1 + \epsilon$ in which $\epsilon \sim N(0, 1)$

Least squares: 2-D example



Complexity of least squares

- If $X_{N \times M}$ and Y_N:
  - Time complexity of OLS is $O(M^2N) + O(M^3) + O(MN) + O(M^2)$. If we omit the lower order items, it leads to: $O(M^2) + O(M^3)$
  - Space complexity of OLS is $O(M^2 + NM)$.

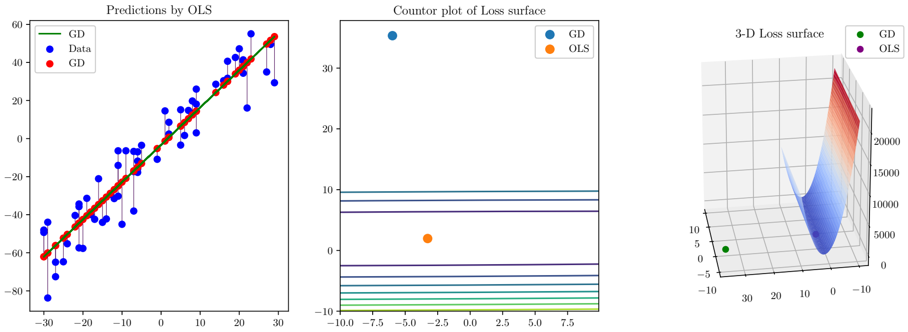
Algorithm 1 Gradient Descent method

1: $x \leftarrow x_0 \in dom(f)$
2: **repeat**
3: $\quad \Delta \mathrm{w} = \text{-} \; \nabla f(x)$
4: $\quad$ Line search: Choose step size t via exact or back-tracing line search
5: $\quad$ Update: $x := x + t\Delta x$
6: **until** $||\nabla f(x)||_2 < \eta = 0$

Algorithm 2 Gradient Descent for linear regression

$x \leftarrow x_0 \in dom(f)$
$m \leftarrow$ Number of training instances
$w_0 \leftarrow$ Initial value
$w_i \leftarrow$ Initial value
$t \leftarrow$ Step
repeat
    $\hat{y} \leftarrow f(x_i, w_i, w_0)$
    $w_0 := w_0 - \frac{t}{m} \sum_{i=1}^{m} (\hat{y} - y_i)^2 \cdot x_0$
    $w_i := w_i - \frac{t}{m} \sum_{i=1}^{m} (\hat{y} - y_i)^2 \cdot x_i$
until $||\Delta f(x)||_2 < \eta = 0$

Least squares vs. Gradient Descent: 2-D example



Logistic Regression

The hypothesis is written as following:

$$h_w(X) = g(W^T X) \tag{7}$$

in which

$$g(z) = \frac{1}{e^{-x} + 1} \tag{8}$$

$$= \begin{cases} Y = 1 & \text{if} \quad g(z) \geq 0.5 \\ Y = 0 & \text{if} \quad g(z) < 0.5 \end{cases} \tag{9}$$

Loss function

The cost function is the binary cross entropy:

$$J(w) = -\frac{1}{m}\sum_{i=1}^{m} y_i \log(h_w(x_i)) + (1 - y_i)\log(1 - h_w(x_i)) \tag{10}$$

- Unlike the case of linear regression, minimizing cross entropy cannot lead to a closed form solution.
- We can only use iterative methods to minimize cross binary entropy.

Algorithm 3 Gradient Descent for logistic regression

$x \leftarrow x_0 \in dom(f)$

$m \leftarrow$ Number of training instances

$w_0 \leftarrow$ Initial value

$w_i \leftarrow$ Initial value

$t \leftarrow$ Step

repeat

    $\hat{y} \leftarrow f(x_i, w_i, w_0)$

    $w_0 := w_0 - \frac{t}{m} \sum_{i=1}^{m} (\hat{y} - y_i)^2 \cdot x_0$

    $w_i := w_i - \frac{t}{m} \sum_{i=1}^{m} (\hat{y} - y_i)^2 \cdot x_i$

until $||\Delta f(x)||_2 < \eta = 0$

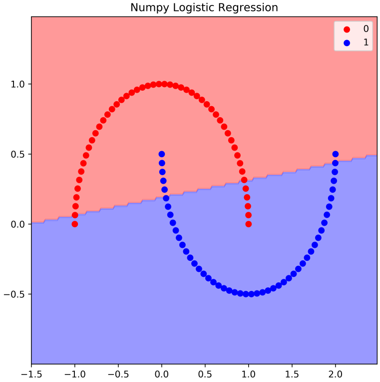
Decision boundary

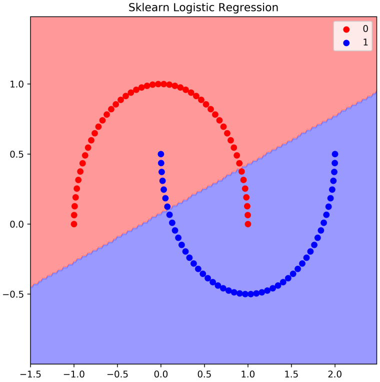
- Decision boundary or decision surface is a hyper-surface that partitions the underlying vector space into two sets, one for each class.
- In statistical classification problems, decision boundary is a hyper-surface in which the predicted label of the class is ambiguous
- If the surface is a hyper-plane, then the classification problem is linear, and the classes are linearly separable

Decision boundary of Logistic Regression

- when $g(Z) >= 0.5$, the predicted label is $Y = 1$. That means when $1 + e^{-W^T X} \leq 2$ and hence $WX \geq 0$.
- Similarly, we can obtain that when $W^T X < 0$, the predicted label is $Y = 0$.

Decision boundary of Logistic Regression





Bias-variance trade-off

In statistical learning theory, the learning of f has two criteria:

- That f can have a minimal empirical risk, namely $R_n(f)$
- But also that f captures an underlying function that extends beyond the sample at hand

$$\mathbb{E}[(y-\hat{f}(x)^2] = \underbrace{\mathbb{E}[\hat{f}(x)] - \mathbb{E}[f(x)]}_{\text{Bias}} + \underbrace{\mathbb{E}[\hat{f}(x)^2] + \mathbb{E}[\hat{f}(x)]^2}_{\text{variance}} + \sigma^2 \tag{11}$$

Regularization

Regularization is a technique one uses to

- prevent over-fitting: the model cannot generalize well to the test/unseen data
- when learning is ill-posed: when there are more features than instances

Theoretically, we say that by using regularization, we are setting *constraints* on the problem we are solving on the class of methods

Regularization methods

$$\text{L1 regularization (Lasso)} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^{n} |w_i| \tag{12}$$

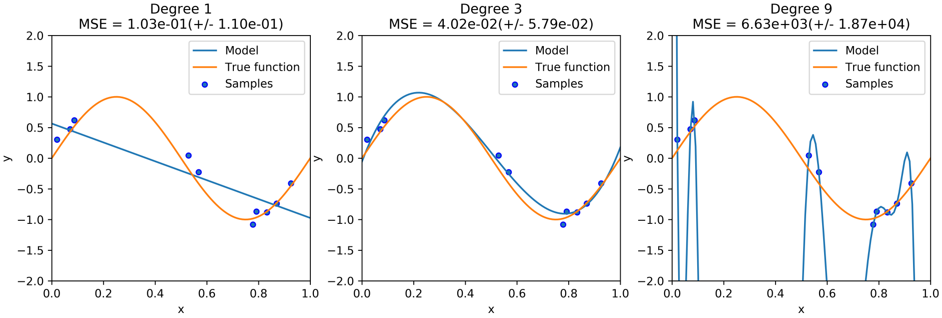
$$\text{L2 regularization (Ridge)} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^{n} w_i^2 \tag{13}$$

$$\text{Elastic net} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda_1 \sum_{i=1}^{n} |w_i| + \lambda_2 \sum_{i=1}^{n} w_i^2 \tag{14}$$

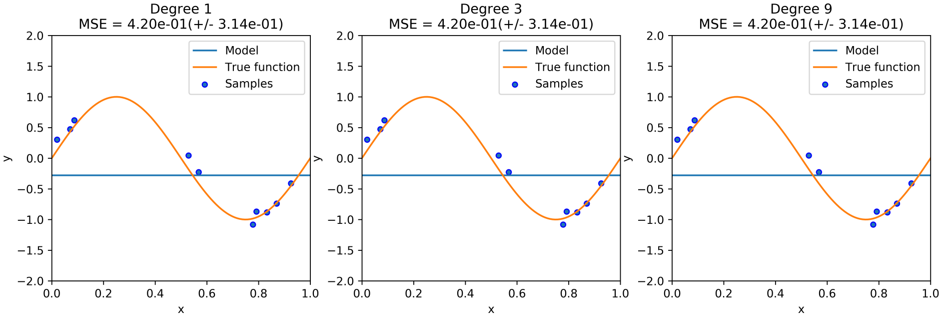
Regularization: example

- Let us show an example where we would like to fit a regression model to $f(x) = \sin(2\pi X)$ with 30 samples.
- For this case, instead of using linear regression in the form $\sum_{i=0}^{n} w_i x_i$ where $x_0 = 0$, we use polynomial features.
- This means that our regression model is in the form of $\sum_{i=0}^{n} w_i x_i^i$ in which n is called the degree of the polynomial and is set by the user of the algorithm.
- Remark: the rationale for choosing polynomial is that it is almost impossible to show the feature of regularization in 2 dimensions!

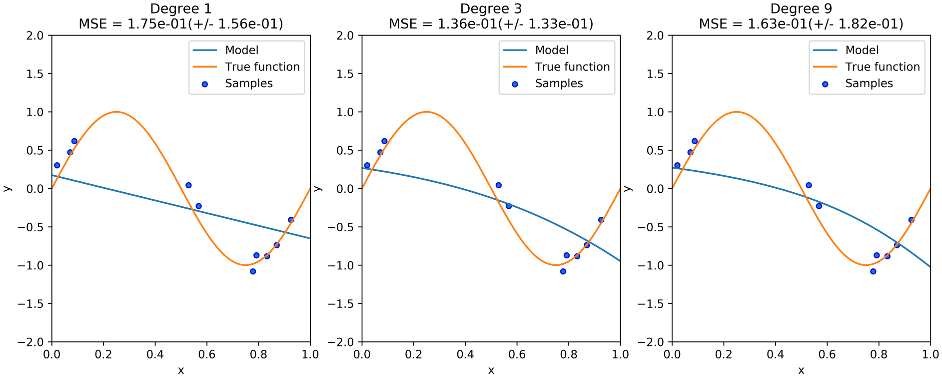
Bias-variance: an example



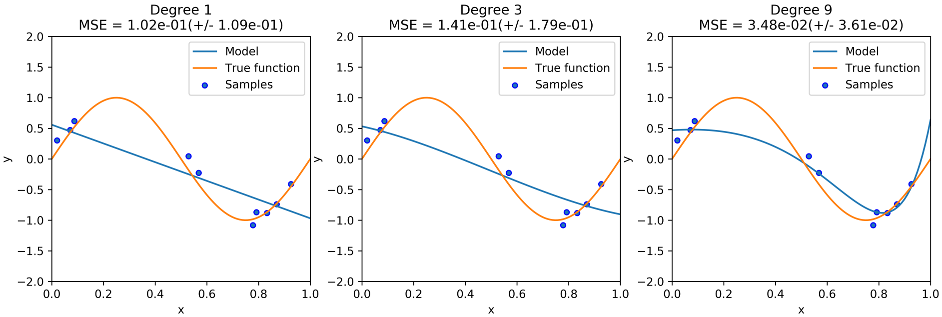
L1 Regularization



L2 Regularization



Elastic Net $\alpha = 10^{-3}$



Regularized Least Squares

One can obtain a closed form version of L2 regularization for Ordinary least squares:

$$\hat{W}_{\text{Ridge}} = (\lambda I_D + X^T X)^{-1}(X^T Y) \tag{15}$$

Algorithm 4 Regularized Gradient Descent for linear models

$x \leftarrow x_0 \in dom(f)$
$m \leftarrow$ Number of training instances
$w_0 \leftarrow$ Initial value
$w_i \leftarrow$ Initial value
$t \leftarrow$ Step
$\lambda \leftarrow$ regularization coefficient
repeat
    $\hat{y} \leftarrow f(x_i, w_i, w_0)$
    $w_0 := w_0 - \frac{t}{m}\sum_{i=1}^{m}(\hat{y} - y_i)^2 \cdot x_0$
    $w_i := w_i - \frac{t}{m}[\sum_{i=1}^{m}(\hat{y} - y_i)^2 \cdot x_i + \lambda w_i]$
until $||\Delta f(x)||_2 < \eta$ =0

Appendix (self-study for students)

Assumptions behind Ordinary Least Squares

- The relationships between dependent variable and the regressors are linear (not a strict assumption)
- Strict exogenity: $E(\epsilon_i|X) = 0 \quad i = 1, ..., N$ with following implications:
  - Unconditional mean of the error is zero, i.e. $\mathbf{E}(\epsilon_i) = 0 \quad i = 1, ..., N$
  - Regressions are orthogonal to the error term for all observations, i.e. $\mathbf{E}(X_j \cdot \epsilon_i) = 0$ for all i, j. As a result, regressors are "contemporaneously" uncorrelated with the error term.
- The rank of the $N \times M$ matrix is M with probability 1 (this means $N \geq M$)
- Spherical error variance, $E(\epsilon_2^2|X) = \sigma^2 > 0$ for $i = 1, ..., N$ and no correlations between observations $E(\epsilon_i \epsilon_j|X) = 0 \quad (i, j = 1, ..., N)$ where $i \neq j$

Measuring the variable importance I

Let us assume the following:

- observations Y_i are uncorrelated with constant variance σ^2
- $X_i \quad i = 1, ..., N$ are fixed (not random anymore)
- Assume that conditional expectation of Y given X is linear:

$$\begin{aligned} Y &= E(Y|X) + \epsilon \\ &= w_0 + \sum_{i=1}^{N} X_i W_i + \epsilon \end{aligned} \tag{16}$$

Estimate the variance, σ^2 by:

$$\hat{\sigma}^2 = \frac{1}{N - M - 1} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{17}$$

Measuring the variable importance II

Then, we can show:

$$\hat{W} \sim \mathcal{N}(W, (\mathbf{XX^T})^{-1}\sigma^2) \tag{18}$$

And then measure the effect of dropping an input variable by calculating Z-score:

$$z_j = \frac{\hat{w}_j}{\hat{\sigma}\sqrt{v_j}} \tag{19}$$

where $v_j = \text{diag}_j(\mathbf{XX^T})^{-1}$

Measuring the variable importance II

Using the calculated z-score, we form a statistical hypothesis test: $H_0 : W_j = 0$ is the null hypothesis. Under the null hypothesis that $W_j = 0$, z_j is distributed as t_{N-m-1}.

Model selection using F-statistic

In addition, we can measure the effect of a group of variables using F-statistic statistical tests. First, we calculate the F-statistic:

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(m_1 - m_0)}{\mathrm{RSS}_1/(N - m_1 - 1)} \tag{20}$$

Where RSS_1 is the RSS of the bigger model with $m_1 + 1$ parameters and RSS_2 has $m_0 + 1$ parameters. After that, we form the null hypothesis that the smaller model is correct. Under the null hypothesis, F-statistic will follow a $F_{m_1 - m_0, N - m_1 - 1}$ statistic. Calculating the CDF of a F distribution using the F-statistic will lead us to a $1 - \alpha$ value. If this value is significant, then the simpler model is correct.

Thanks!