Machine Learning: Algorithms and Applications

Advanced Multimedia Research Lab
University of Wollongong

Regression

Outline

1. Introduction - historical and theoretical
2. Linear Regression
3. Kernel Ridge regression
4. Lasso Regression

Regression - Historical and general ideas

- Regression is a supervised learning method often used in prediction tasks (with modification, also in classification)
- Regression as a scientific method first appeared around 1885
- Francis Galton developed the ideas in the studies of heredity stature - comparison of height of parents and their children (Izenman 2008)
- Galton did not link the least squares method and regression which was discovered 80 years later
- George Yule (1897) showed that an assumption of a Gaussian error in regression could be replaced by assumption that variables are linearly related - hence least squares can be applied to regression
- Linear regression models can be simple, multiple or multivariate
  1. simple linear regression - one input and one output
  2. multiple regression - many inputs and one output
  3. multivariate regression - many inputs and many outputs
- In general there is the output (also called the dependent variable) that is assumed to be linearly related to the input(s) (also called the independent variables; input space)
- Independent variables could be formed from a linear combination of a fixed set of nonlinear functions (basis functions) of input variables
- It is the coefficients of the function of relatedness that we want to determine and obtain an equation for use in prediction on new observed variables

Regression - Theoretical development

Regression problem

- Let $\mathcal{X}$ denote the input space and $\mathcal{Y}$ a measurable subset of $\mathbb{R}$.
- Denote by $\mathcal{D}$ an unknown distribution over $\mathcal{X}$ according to which the inputs are drawn
- Let $f : \mathcal{X} \to \mathcal{Y}$ be the target labelling function
- This is a deterministic learning scenario; a stochastic learning scenario will have distribution over pairs $(x, y) \in \mathcal{X} \times \mathcal{Y}$
- Learner receives a labelled sample $S = \{(x_1, y_1), \dots, (x_m, y_m)\} \in (\mathcal{X} \times \mathcal{Y})$ with $x_1, \dots, x_m$ drawn i.i.d from $\mathcal{D}$ and $y_i = f(x_i)$ for all $i \in [1, m]$.
- Denote by $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ the loss function measuring the magnitude of error
  - Commonly, squared error is used: $\mathcal{L}(y, \hat{y}) = |y - \hat{y}|^2$ for all $y, \hat{y} \in \mathcal{Y}$
  - Generally, $\mathcal{L}_p$ loss function may be used: $\mathcal{L}(y, \hat{y}) = |y - \hat{y}|^p$ for all $y, \hat{y} \in \mathcal{Y}$ and some $p \geq 1$
- Given a hypothesis set $\mathcal{H}$ of functions mapping $\mathcal{X}$ to $\mathcal{Y}$, regression problem consists of using the labelled sample $\mathcal{S}$ to find the hypothesis $h \in \mathcal{H}$ with small expected loss or generalization error $\mathcal{R}(h)$ with respect to the target function, f:

$$\mathcal{R}(h) = E_{x \sim \mathcal{D}}[\mathcal{L}(h(x), f(x))] \tag{1}$$

- Empirical loss is:

$$\hat{\mathcal{R}}(h) = \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(h(x_i), y_i) \tag{2}$$

Quick note on generalization bounds

- If loss function $\mathcal{L}$ is bounded by some $M > 0$ it results in a bounded regression problem; i.e.:
  - $\mathcal{L}(y, \hat{y}) \leq M$ for all $y, \hat{y} \in \mathcal{Y}$;
  - more strictly $\mathcal{L}(h(x), f(x)) \leq M$ for all $h \in \mathcal{H}$ and $x \in \mathcal{X}$
- Without proof we state the following theorem on generalization bound for regression problem:

Theorem (Regression generalization bound)

Let $\mathcal{L}$ be a bounded loss function. Assume that the hypothesis set $\mathcal{H}$ is finite. Then , for $\delta > 0$, with probability at least $1 - \delta$, the following inequality holds for all $h \in \mathcal{H}$:

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + M\sqrt{\frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{2m}} \tag{3}$$

- The theorem above indicates that the empirical and generalization errors are made as close as possible by making $M\sqrt{\frac{\log|\mathcal{H}|+\log\frac{1}{\delta}}{2m}}$ as small as possible
- As an exercise, explore how the cardinality of hypothesis set $\mathcal{H}$ ($|\mathcal{H}|$), the number δ, the bound on the loss function, M, and the number of training samples, m, individually affects the generalization error. Hint: keep some values constant and explore the effect of varying one variable

Linear regression

- Let $\mathbf{\Phi} : \mathcal{X} \to \mathbb{R}^N$ be a feature mapping from input space $\mathcal{X}$ to $\mathbb{R}^N$
- Consider a family of linear hypotheses

$$\mathcal{H} = \{x \mapsto \mathbf{w}.\mathbf{\Phi}(x) + b : \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}\} \tag{4}$$

- Linear regression seeks an hypothesis in $\mathcal{H}$ with the smallest mean squared error
- Given a sample set $\mathcal{S} = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$ we need to solve the following optimization problem:

$$\min_{w,b} \frac{1}{m} \sum_{i=1}^{m} (\mathbf{w}.\mathbf{\Phi}(x_i) + b - y_i)^2 \tag{5}$$

- If we write $\boldsymbol{X} = \begin{bmatrix} \Phi(x_1) & \ldots & \Phi(x_m) \\ 1 & \ldots & 1 \end{bmatrix}$, $\boldsymbol{W} = \begin{bmatrix} w_1 \\ \vdots \\ w_N \\ 1 \end{bmatrix}$ and $\boldsymbol{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$ the optimization problem in (5) can be written compactly as

$$\min_{\boldsymbol{W}} F(\boldsymbol{W}) = \frac{1}{m} ||\boldsymbol{X}^T \boldsymbol{W} - \boldsymbol{Y}||^2 \tag{6}$$

Linear regression

- Consider the dimensions of the entries in Equation (6)
  - $\boldsymbol{X}^T \in \mathbb{R}^{m\times(N+1)}$
  - $\boldsymbol{W} \in \mathbb{R}^{N+1}$
  - $\boldsymbol{X}^T\boldsymbol{W} \in \mathbb{R}^{m}$
  - $\boldsymbol{Y} \in \mathbb{R}^{m}$
- In transforming Equation (5) to Equation (6) we have done the following:

$$\begin{aligned} y_i &= w_i x_i + b \\ &= w_i' x_i + 1 \end{aligned}$$

where the bias b has been absorbed in the weight w'

- The optimization problem in Equation (6), $F(\boldsymbol{W})$, is convex, differentiable and has a global minimum that can be obtained by differentiating $F(\boldsymbol{W}) = \frac{1}{m}||\boldsymbol{X}^T\boldsymbol{W} - \boldsymbol{Y}||^2$ with respect to $\boldsymbol{W}$ and equating to zero
- $\nabla F(\boldsymbol{W}) = 0$; $\frac{2}{m}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W} - \boldsymbol{Y}) = 0$ from which $\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{W} = \boldsymbol{X}\boldsymbol{Y}$

$$\boldsymbol{W} = \begin{cases} (\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\boldsymbol{Y} & \text{if } \boldsymbol{X}\boldsymbol{X}^T \text{ is invertible} \\ (\boldsymbol{X}\boldsymbol{X}^T)^{\dagger}\boldsymbol{X}\boldsymbol{Y} & \text{otherwise; using the pseudo-inverse } \dagger \end{cases} \tag{7}$$

Linear Regression

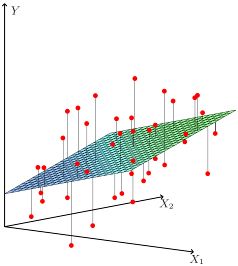


Figure 1: Linear least square fitting ($X \in \mathbb{R}^2$). We seek the linear function of X that minimizes sum of squared errors from Y (Hastie et al. 2001).

Linear Regression

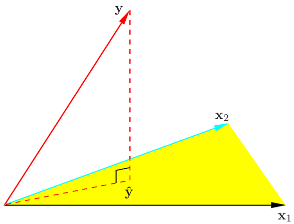


Figure 2: N-dimensional geometry of least squares regression with two independent variables $\boldsymbol{x_1}, \boldsymbol{x_2}$. Predicted y vector is orthogonally projected onto the hyperplane spanned by $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$. $\hat{y}$ represents the vector of the least squares predictions (Hastie et al. 2001).

Linear Regression

- Results shown in Equation (7) is also referred to as the least squares estimate of the weight vector (coefficients), ***W***, of the linear regression model
- Important notes on linear regression:
  - Prediction accuracy of least squares estimate often has low bias but large variance[1]
  - If there are a large number of independent variables it is desirable to know the key variables that exhibit strong effect
  - There is no strong generalization guarantee because we only minimize empirical error without controlling the norm (length) of the weight vector; there is no regularization

[1] See Figure (4) ▸ here

Logistic Regression

- In linear regression, the outcome variable is a continuous variable.
- When the outcome variable is categorical in nature, logistic regression can be used
  - To predict the probability of an outcome based on the input variables.

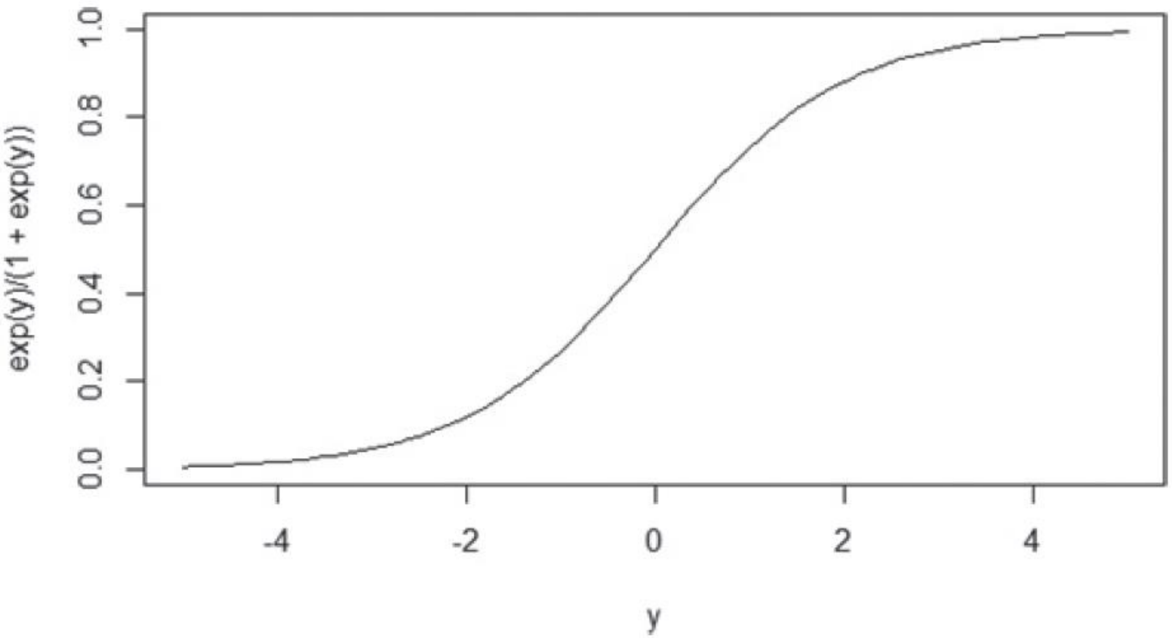
Logistic Regression

- Use Cases
  - Medical: determine the probability of a patient’s response to a medical treatment.
  - Finance: determine the probability that an applicant will default on the loan.
  - Marketing: Determine the probability for a customer to switch carriers (churning).
  - Engineering: Determine the probability of a mechanical part to fail.

Model Description

- Logistic regression is based on the logistic function:

$$f(y) = \frac{e^y}{1+e^y} \qquad \text{for } -\infty < y < \infty$$



Model Description

- In logistic regression, *y* is expressed as a linear function of the input variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_{p-1} x_{p-1}$$

- Then the probability of an event is computed:

$$p(x_{1,} x_{2,} \ldots, x_{p-1}) = f\left(y\right) = \frac{e^y}{1 + e^y} \quad \text{for} -\infty < y < \infty$$

- Note: Only f(y) is observed, not y.

Model Description

- Rewriting the equation can give us the log odd ratio (the logit of *p*)

$$ln\left(\frac{p}{1-p}\right) = y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_p x_{p-1}$$

- Maximum Likelihood Estimation (MLE) is often used to estimate the model parameters
  - It finds the parameter values that maximize the chances of observing the given dataset.

Kernel Ridge regression

- Formulation is somewhat similar linear regression; consider mapping from input space to a feature space but with a kernel $\Phi(\cdot)$

- This model gives better theoretical guarantees and improved performance in practice (there is a theorem that supports this claim) The optimization problem is written compactly as:

  $$\min_{\boldsymbol{W}} F(\boldsymbol{W}) = \lambda||\boldsymbol{W}||^2 + ||\boldsymbol{X}^T\boldsymbol{W} - \boldsymbol{Y}||^2 \tag{8}$$

  where λ is a positive parameter that determines the trade-off between the regularization term $||\boldsymbol{W}||^2$ and the empirical mean squared error; $\boldsymbol{X} \in \mathbb{R}^{N\times m}$ is the matrix of feature vectors, $\boldsymbol{X} = [\Phi(x_1), \ldots, \Phi(x_m)]$ and $\boldsymbol{W}$ and $\boldsymbol{Y}$ are as defined previously (see Equation (6))

- Optimization problem of Equation (8) is convex and differentiable with a global minimum if and only if

  $$\nabla F(\boldsymbol{W}) = 0 \Leftrightarrow (\boldsymbol{X}\boldsymbol{X}^T + \lambda\boldsymbol{I})\boldsymbol{W} = \boldsymbol{X}\boldsymbol{Y} \Leftrightarrow \boldsymbol{W} = (\boldsymbol{X}\boldsymbol{X}^T + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}\boldsymbol{Y} \tag{9}$$

  $\boldsymbol{X}\boldsymbol{X}^T + \lambda\boldsymbol{I}$ is always invertible [2]

- Alternative formulation of the kernel ridge regression

  $$\min_{\boldsymbol{w}} \sum_{1}^{m} (\boldsymbol{w}\cdot\boldsymbol{\Phi}(x_i) - y_i)^2 \qquad \text{subject to} \qquad ||\boldsymbol{w}||^2 \leq \Lambda^2 \tag{10}$$

[2] because its eigenvalues are sum of non-negative eigenvalues of positive semi-definite matrix $\boldsymbol{X}\boldsymbol{X}^T$ and $\lambda > 0$

Kernel Ridge regression

Some properties of ridge regression:

- In essence it is a model selection method in which the ridge parameter λ helps select/weight the variables appropriately.
- The choice of the ridge parameter is a tool to balance the “bias-variance” trade-off. The larger the value of λ the larger the bias and the smaller the variance. The parameter can be determined using cross validation technique.
- The ridge regression estimator is a shrinkage estimator that shrinks the least square weights toward zero.
- It can be used with (positive definite symmetric PDS) kernels and hence can be extended to non-linear regression and more general feature spaces.

Lasso Regression

- Our goal in prediction is to choose an economical (parsimonious) model that will balance the bias-variance trade-off.

- What variables are important for the prediction?

- Variable selection is another method of solving this problem
  1. Backward elimination: Begin with full set of variables and drop at each step the variable whose $F-$ratio is smallest:

     $$F = \frac{(RSS_0 - RSS_1)/(df_0 - df_1)}{RSS_1/df_1} \tag{11}$$

     $RSS_0 = \sum_i (y_i - \hat{y}_i)^2$ computed with reduced model and with degree of freedom df_0 ;
     $RSS_1 = \sum_i (y_i - \hat{y}_i)^2$ computed with larger model and with degree of freedom df_1 ;
     The reduced model is refitted and the iteration is repeated.
  2. Forward selection: Begin with an empty set of variables and select the variable from the list that gives the largest F value[3].

[3] More on feature selection later in the lecture series.

Lasso Regression

- Lasso is a short for Least absolute shrinkage and selection operator
- Essentially it combines variable subset selection and shrinkage to improve accuracy
- This model does not allow an easy use of a PDS kernel; assume input space $\mathcal{X}$, is a subset of $\mathbb{R}^N$
- Consider a family of linear hypotheses

$$\mathcal{H} = \{x \mapsto \mathbf{w}.\boldsymbol{x} + b : \boldsymbol{w} \in \mathbb{R}^N, b \in \mathbb{R}\} \tag{12}$$

- Given a sample set $\mathcal{S} = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$
- Lasso regression seeks an hypothesis in $\mathcal{H}$ that minimizes empirical squared error with a regularization term depending on the norm of the weight vector;
- Lasso uses L_1 norm instead of L_2 norm (ridge regression - see Equations (8) and (10)):

$$\min_{\boldsymbol{w},b} F(\boldsymbol{w}, b) = \lambda||\boldsymbol{w}||_1 + \sum_{i=1}^{m} (\mathbf{w}.\boldsymbol{x}_i + b - y_i)^2 \tag{13}$$

- Equivalently:
  $\min_{\boldsymbol{w},b} \sum_{i=1}^{m} (\mathbf{w}.\boldsymbol{x}_i + b - y_i)^2$ subject to $||\boldsymbol{w}||_1 \leq \Lambda_1$; It is a Quadratic Program solvable by QP solvers

Lasso Regression

- Key property of Lasso is that it leads to sparse solution of $\boldsymbol{w}$ - one with few non-zero components
- Sparsity is encouraged by L_1 norm

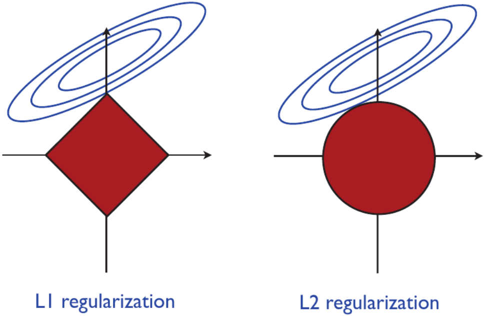


Figure 3: Comparision of Lasso and ridge regression solutions (Mohri et al. 2012)

- Objective function is quadratic and contours are ellipsoids (See Figure 3); Lasso solution is intersection with L_1 ball occurring at corner where some coordinates are zero, hence it promotes sparsity; contrast with L_2 regularization

Model Selection and variance-bias Trade-off

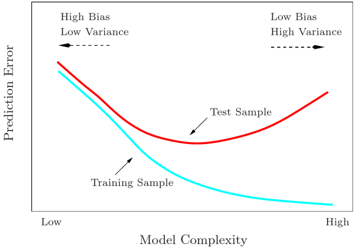


Figure 4: Typical training and test error behaviour as a function of model complexity (Hastie et al. 2001). Training error decreases as model complexity increases; model overfits leading to poor generalization and large variance. Test error increases if model is not complex enough; model underfits; lead to large bias and poor generalization. So there is a bias-variance trade-off.

- The prediction error has three parts:
  1. irreducible error (variance of the new test target) which is beyond our control
  2. Bias component - the squared difference between true mean of the estimate and the expected value of the estimate
  3. Variance component - variance of an average

Back to ▸ main

Bibliography

Alpaydin, E. (2010), *Introduction to Machine Learning*, second edn, The MIT Press, Cambridge Massachusetts.

Duda, R. O., Hart, P. E. & Stork, D. G. (2001), *Pattern Classification*, Second edn, John Wiley and Sons.

Goodfellow, I., Bengio, Y. & Courville, A. (2016), *Deep Learning*, MIT Press.

Hastie, T., Tibshirani, R. & Friedman, J. (2001), *The Elements of Statistical Learning - Data Mining, Inference and Prediction*, Springer Science+Business Media LLC.

Izenman, A. J. (2008), *Modern Multivariate Statistical Techniques - Regression, Classification and Manifold Learning*, Springer Science+Business Media LLC.

Kellleher, J. D., Namee, B. M. & D'Arcy, A. (2015), *Fundamentals of Machine Learning for Predictive Data Analytics - Algorithms, Worked Examples and Case Studies*, The MIT Press, Cambridge Massachusetts.

Mitchell, T. M. (1997), *Machine Learning*, WCB McGraw-Hill.

Mohri, M., Rostamizadeh, A. & Talwalkar, A. (2012), *Foundations of Machine Learning*, MIT Press.

Webb, A. (2002), *Statistical Pattern Recognition*, Second edn, John Wiley and Sons.