

MACHINE LEARNING

In Q1 to Q11, only one option is correct, choose the correct option:

1. Which of the following methods do we use to find the best fit line for data in Linear Regression?
☒ A) Least Square Error B) Maximum Likelihood
C) Logarithmic Loss D) Both A and B
2. Which of the following statement is true about outliers in linear regression?
☒ A) Linear regression is sensitive to outliers B) linear regression is not sensitive to outliers
C) Can't say D) none of these
3. A line falls from left to right if a slope is _____?
A) Positive ☒ B) Negative
C) Zero D) Undefined
4. Which of the following will have symmetric relation between dependent variable and independent variable?
A) Regression ☒ B) Correlation
C) Both of them D) None of these
5. Which of the following is the reason for over fitting condition?
A) High bias and high variance B) Low bias and low variance
☒ C) Low bias and high variance D) none of these
6. If output involves label then that model is called as:
A) Descriptive model ☒ B) Predictive modal
C) Reinforcement learning D) All of the above
7. Lasso and Ridge regression techniques belong to _____?
A) Cross validation B) Removing outliers
C) SMOTE ☒ D) Regularization
8. To overcome with imbalance dataset which technique can be used?
☒ A) Cross validation B) Regularization
C) Kernel D) SMOTE
9. The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary classification problems. It uses _____ to make graph?
A) TPR and FPR B) Sensitivity and precision
☒ C) Sensitivity and Specificity D) Recall and precision
10. In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the curve should be less.
A) True ☒ B) False
11. Pick the feature extraction from below:
☒ A) Construction bag of words from a email
☒ B) Apply PCA to project high dimensional data
C) Removing stop words
D) Forward selection

In Q12, more than one options are correct, choose all the correct options:

12. Which of the following is true about Normal Equation used to compute the coefficient of the Linear Regression?
☒ A) We don't have to choose the learning rate.
☒ B) It becomes slow when number of features is very large.
☒ C) We need to iterate.
D) It does not make use of dependent variable.
-

Q13 and Q15 are subjective answer type questions, Answer them briefly.

1. Explain the term regularization?

Regularization is a approach utilized in regression to lessen the complexity of the version and to cut back the coefficients of the impartial features. This method lowers the coefficients for less computing expense while reducing the possibility of overfitting by simplifying a complicated model.

Standard least-squares models typically have some variance, meaning they don't transfer well to data sets other than their training sets. Regularization greatly lowers the model's variance while maintaining or even increasing its bias.

Therefore, the impact on bias and variance is controlled by the regularisation value λ , which is used in the strategies discussed above. As the value of λ increases, the coefficient values decrease, lowering the variance. This increase λ in is advantageous because it just reduces variance (avoiding overfitting), without affecting any significant characteristics of the data. But after a certain point, the model begins to lose crucial characteristics, leading to bias in the model and underfitting the data. Therefore, it is important to choose value of λ wisely.

2. Which particular algorithms are used for regularization?

Particularly there are three types of algorithms are used for regularization

1. Ridge Regression
2. LASSO (Least Absolute Shrinkage and Selection Operator) Regression
3. Elastic-Net Regression

Ridge Regression

Ridge regression is a methodology for analyzing information that suffer from multi-collinearity.

$$Loss = \sum_{i=1}^n (y_i - (w_i x_i + c))^2 + \lambda \sum_{i=1}^n w_i^2$$

Loss Function for Ridge Regression

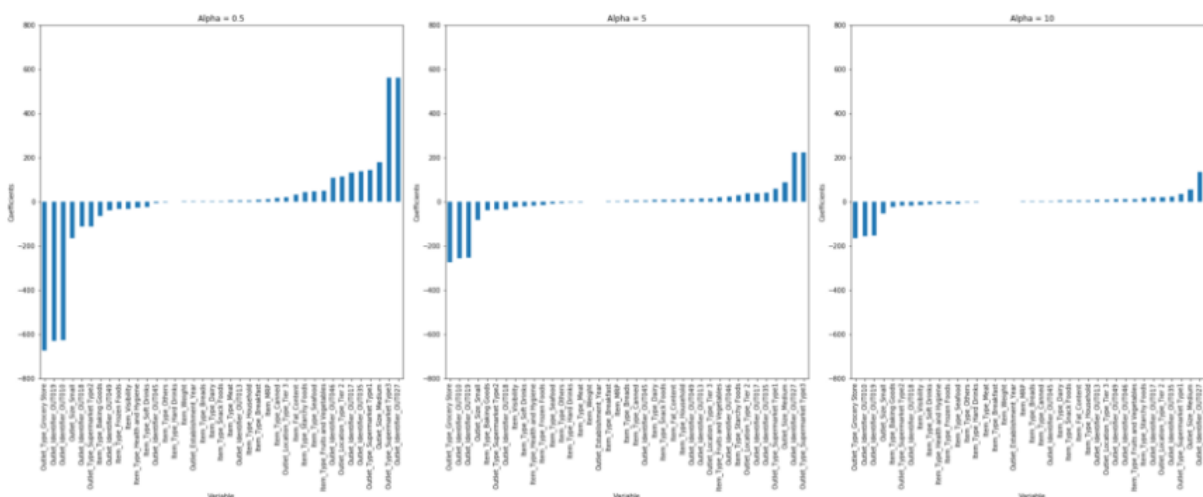
regression adds a penalty (L2 penalty) to the loss function that is equivalent to the square of the magnitude of the coefficients.

The regularization parameter (λ) regularizes the coefficients such that if the coefficients take large values, the loss function is penalized.

- $\lambda \rightarrow 0$, the penalty term has no effect, and the estimates produced by ridge regression will be equal to least-squares i.e. the loss function resembles the loss function of the Linear Regression algorithm. Hence, a lower value of λ will resemble a model close to the Linear regression model.
- $\lambda \rightarrow \infty$, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero (coefficients are close to zero, but not zero).

Note: Ridge regression is also known as the L2 Regularization.

To sum up, Ridge regression shrinks the coefficients as it helps to reduce the model complexity and multi-collinearity.



Ridge Regression: Coefficient values if $\lambda = 0.5, 5$ and 10 respectively || [Source](#)

LASSO Regression

LASSO is a regression analysis method that performs both feature selection and regularization in order to enhance the prediction accuracy of the model.

$$Loss = \sum_{i=1}^n (y_i - (w_i x_i + c))^2 + \lambda \sum_{i=1}^n |w_i|$$

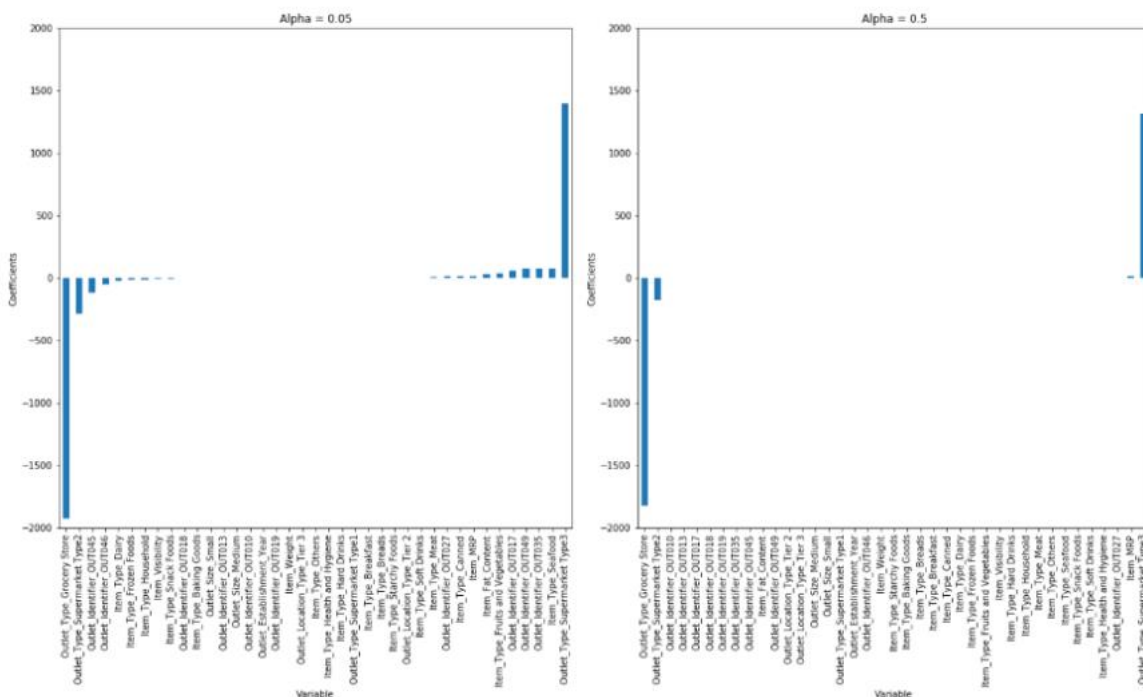
Loss Function for LASSO Regression

LASSO regression adds a penalty (L1 penalty) to the loss function that is equivalent to the magnitude of the coefficients.

In LASSO regression, the penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the regularization parameter λ is sufficiently large.

Note: LASSO regression is also known as the L1 Regularization (L1 penalty).

To sum up, LASSO regression converts coefficients of less important features to zero, which indeed helps in feature selection, and it shrinks the coefficients of remaining features to reduce the model complexity, hence avoiding overfitting.



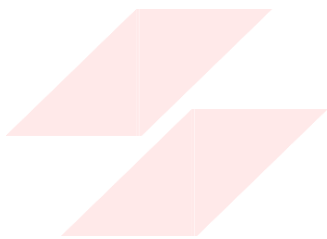
LASSO Regression: Coefficient values if $\lambda = 0.05$, and 0.5 respectively || [Source](#)

Elastic-Net Regression

Elastic Net is a regularized regression model that combines ℓ_1 and ℓ_2 penalties, i.e., lasso and ridge regression. We have discussed the limitations of lasso regression, where we found the incapability of lasso is choosing the number of predictors. The elastic net includes the penalty of lasso regression, and when used in isolation, it becomes the ridge regression.

$$Loss = \sum_{i=0}^n (y_i - (w_i x_i + c))^2 + \lambda_1 \sum_{i=0}^n |w_i| + \lambda_2 \sum_{i=0}^n w_i^2$$

Loss Function for Elastic-Net Regression



FLIP ROBO

MACHINE LEARNING

15. Explain the term error present in linear regression equation?

It is often said that the error term in a regression equation represents the effect of the variables that were omitted from the equation. This is unsatisfactory, even in simple contexts, as the following discussion should indicate. Suppose subjects are IID, and all variables are jointly normal with expectation 0. Suppose the explanatory variables have variance 1. The explanatory variables may be correlated amongst themselves, but any p of them have a non-singular p -dimensional distribution. The parameters α_j are real. Let

$$(1) \quad Y_i = \sum_{j=1}^{\infty} \alpha_j X_{ij}$$

For each $p = 1, 2, \dots$, consider the regression model

$$(2) \quad Y_i = \sum_{j=1}^p \alpha_j X_{ij} + \epsilon_i(p)$$

where

$$(3) \quad \epsilon_i(p) = \sum_{j=p+1}^{\infty} \alpha_j X_{ij}$$

The α_j are identifiable. If the X_{ij} are independent for $j = 1, 2, \dots$, the standard assumptions hold, and $\epsilon_i(p)$ does indeed represent the effect on Y_i of the omitted variables $\{X_{ij} : j = p+1, \dots\}$, at least in an algebraic sense. On the other hand, if the X_{ij} are dependent, the matter is problematic. If we take (1–3) as written, then $\epsilon_i(p)$ represents the effect on Y_i of the omitted variables—but $\epsilon_i(p)$ is correlated with the explanatory variables. The standard assumptions fail, and fitting (2) to data for $i = 1, \dots, n$ will estimate the wrong parameters. If $\epsilon_i(p)$ is replaced by $\epsilon_i(p)^\perp$, namely, the part of $\epsilon_i(p)$ independent of X_{i1}, \dots, X_{ip} , we have a bona fide regression model, but with different α 's.

There is no easy way out of the difficulty. The conventional interpretation for error terms needs to be reconsidered. At a minimum, something like this would need to be said: the error term represents the combined effect of the omitted variables, assuming that

- (i) the combined effect of the omitted variables is independent of each variable included in the equation,
- (ii) the combined effect of the omitted variables is independent across subjects,
- (iii) the combined effect of the omitted variables has expectation 0.

This is distinctly harder to swallow. Pratt and Schlaifer have a discussion in great depth.

MACHINE LEARNING

Some technical details

If the α_j vanish for all but finitely many j , there are no technical issues. The inferential issue remains, provided the largest j with $\alpha_j \neq 0$ is an unknown parameter. Suppose next that $\alpha_j \neq 0$ for infinitely many j . Summability and identifiability must be demonstrated. To avoid interesting but unnecessary probabilistic complications, suppose $\sum_j |\alpha_j| < \infty$. Fix i . Suppose also that part of each $X_{ij} : j = 1, 2, \dots$ is independent of all the other X_{ik} , and has L_2 norm at least $\eta > 0$. More specifically, let X_{ij}^\perp be X_{ij} net of $\{X_{ik} : k = 1, \dots, p \text{ with } k \neq j\}$. Thus, we assume $\|X_{ij}^\perp\| \geq \eta$, where $\|\cdot\|$ is the L_2 norm. See below for definitions and some theory.

Now $\|\epsilon_i(p)\| \leq \sum_{j=p+1}^\infty |\alpha_j|$ is small, so the sum on the right hand side of (1) converges in L_2 . Fix j and p with $1 \leq j \leq p$. The regression of $\epsilon_i(p)$ on $\{X_{i1}, \dots, X_{ip}\}$ has a small coefficient on X_{ij} , because

- (i) $\epsilon_i(p)$ is small,
- (ii) we get the coefficient by regressing $\epsilon_i(p)$ on X_{ij}^\perp , and
- (iii) $\|X_{ij}^\perp\| \geq \eta$.

In more formal terms, by Lemma 2 below, a regression of Y_i on X_{i1}, \dots, X_{ip} in the random-variable domain gives a coefficient on X_{ij} of $\text{cov}(X_{ij}^\perp, Y)/\text{var}(X_{ij}^\perp)$. This coefficient is α_j , with an error that is at most

$$(4) \quad \frac{\text{cov}(X_{ij}^\perp, \epsilon_i(p))}{\text{var}(X_{ij}^\perp)} \leq \frac{\|X_{ij}^\perp\| \|\epsilon_i(p)\|}{\|X_{ij}^\perp\|^2} \leq \eta^{-1} \|\epsilon_i(p)\| \leq \eta^{-1} \sum_{j=p+1}^\infty |\alpha_j| \rightarrow 0$$

as $p \rightarrow \infty$. That proves identifiability.

A mistake to avoid

Some may conclude from the forgoing that bigger models are better. Perhaps, but (i) eventually we run out of data, and (ii) there is always the ugly possibility of inadvertently including an endogenous variable. Also see exercise 15 on page 105 of *Statistical Models* for information on standard errors in the presence of misspecification. Kitchen-sink models have their problems too.

Regression in the domain of random variables

Changing notation, let q be a positive integer. Let U_1, \dots, U_q, V be jointly normal random variables, each having expectation 0. Let $C_{ij} = \text{cov}(U_i, U_j)$. This is a symmetric $q \times q$ matrix, assumed to be positive definite. Let $D_i = \text{cov}(U_i, V)$. Take $D = (D_1, \dots, D_q)'$ as a $q \times 1$ vector. Let $B = C^{-1}D$, which is also a $q \times 1$ vector. Let $V^\perp = V - (U_1, \dots, U_q) \times B$, a scalar random variable.

Lemma 1. (i) V^\perp is normal with expectation 0, and (ii) $V^\perp \perp (U_1, \dots, U_q)$ in the sense that $\text{cov}(U_j, V^\perp) = E(U_j V^\perp) = 0$ for each $j = 1, \dots, q$. In particular, (iii) V^\perp and (U_1, \dots, U_q) are independent.

For the proof, assertion (i) is immediate. For (ii), we need only check that

$$\text{cov}(U_j, V) = \text{cov}(U_j, (U_1, \dots, U_q) \times B) = \sum_{k=1}^q \text{cov}(U_j, U_k) B_k = \sum_{i=1}^q C_{jk} B_k,$$

MACHINE LEARNING

i.e., $D = CB$. But $B = C^{-1}D$ by construction, completing the proof.

In short, $(U_1, \dots, U_q) \times B$ is the regression of V on U_1, \dots, U_q ; the coefficient on U_i is B_i ; and V^\perp is the part of V independent of U_1, \dots, U_q . This is also “ V net of U_1, \dots, U_q .” Normality is relevant only to convert orthogonality into independence. Without normality, $(U_1, \dots, U_q) \times B$ is the linear projection of V onto U_1, \dots, U_q , i.e., the linear combination of U_1, \dots, U_q closest to V in L_2 —because V^\perp is orthogonal to U_1, \dots, U_q . The simplest special case has $q = 1$. Then the regression coefficient takes a form that may be more familiar, $\text{cov}(U_1, V)/\text{var}(U_1)$.

Lemma 2. The regression of V on $U = (U_1, \dots, U_q)$ can be computed by the following stepwise procedure, with $\tilde{U} = (U_2, \dots, U_q)$.

- (i) Regress V on U_2, \dots, U_q . Let α be the $(q - 1) \times 1$ vector of regression coefficients. Let $\hat{V} = \tilde{U}\alpha$ and $V^\perp = V - \hat{V}$.
- (ii) Regress U_1 on U_2, \dots, U_q . Let β be the $(q - 1) \times 1$ vector of regression coefficients. Let $\hat{U}_1 = \tilde{U}\beta$ and $U_1^\perp = U_1 - \hat{U}_1$.
- (iii) Regress V on U_1^\perp . Let γ be the regression coefficient, a scalar.

The $q \times 1$ vector of regression coefficients of V on U_1, \dots, U_q is then



$$\begin{pmatrix} \gamma \\ \alpha - \beta\gamma \end{pmatrix}$$

Proof. Since $V = \hat{V} + V^\perp$ and $\hat{V} \perp U_1^\perp$, whether we regress V on U_1^\perp or V^\perp on U_1^\perp , the coefficient on U_1^\perp will be the same, viz., γ . So $\epsilon = V^\perp - U_1^\perp\gamma \perp U_1^\perp$. Plainly, $\epsilon \perp U_2, \dots, U_q$, because ϵ is a linear combination of V^\perp and U_1^\perp . Thus,

$$\begin{aligned}
 (5) \quad V &= \hat{V} + V^\perp \\
 &= \hat{V} + U_1^\perp\gamma + \epsilon \\
 &= \tilde{U}\alpha + (U_1 - \tilde{U}\beta)\gamma + \epsilon \\
 &= U_1\gamma + \tilde{U}(\alpha - \beta\gamma) \\
 &= \begin{pmatrix} \gamma \\ \alpha - \beta\gamma \end{pmatrix} U + \epsilon
 \end{aligned}$$

with $\epsilon \perp U$, as required. To clarify the notation, U is $1 \times q$ and \tilde{U} is $1 \times (q - 1)$; both are random vectors; \hat{V} , V^\perp , \hat{U}_1 , U_1^\perp , ϵ are all scalar random variables. If U_1, \dots, U_q , V are taken as jointly normal, these derived quantities are jointly normal too. The quantities α, β, γ are parameters not estimates, being computed from the joint distribution not from data. Exercise 17 on page 34 of *Statistical Models* covers regression in the data domain using a method exactly like that in Lemma 2, although the notation is little different.