Ridge\_regression.py

# \*\*Ridge regression\*\* (or "L2 regularization") minimizes: $$\text{RSS} + \alpha \sum\_{j=1}^p \beta\_j^2$$

#

# \*\*Lasso regression\*\* (or "L1 regularization") minimizes: $$\text{RSS} + \alpha \sum\_{j=1}^p |\beta\_j|$$

#

# - $p$ is the \*\*number of features\*\*

# - $\beta\_j$ is a \*\*model coefficient\*\*

# - $\alpha$ is a \*\*tuning parameter:\*\*

# - A tiny $\alpha$ imposes no penalty on the coefficient size, and is equivalent to a normal linear regression model.

# - Increasing the $\alpha$ penalizes the coefficients and thus shrinks them.

# ### Lasso and ridge path diagrams

#

# A larger alpha (towards the left of each diagram) results in more regularization:

#

# - \*\*Lasso regression\*\* shrinks coefficients all the way to zero, thus removing them from the model

# - \*\*Ridge regression\*\* shrinks coefficients toward zero, but they rarely reach zero

#

# Source code for the diagrams: [Lasso regression](http://scikit-learn.org/stable/auto\_examples/linear\_model/plot\_lasso\_lars.html) and [Ridge regression](http://scikit-learn.org/stable/auto\_examples/linear\_model/plot\_ridge\_path.html)

# ![Lasso and Ridge Path Diagrams](images/lasso\_ridge\_path.png)

# ### Advice for applying regularization

#

# \*\*Should features be standardized?\*\*

#

# - Yes, because otherwise, features would be penalized simply because of their scale.

# - Also, standardizing avoids penalizing the intercept, which wouldn't make intuitive sense.

#

# \*\*How should you choose between Lasso regression and Ridge regression?\*\*

#

# - Lasso regression is preferred if we believe many features are irrelevant or if we prefer a sparse model.

# - If model performance is your primary concern, it is best to try both.

# - ElasticNet regression is a combination of lasso regression and ridge Regression.

# ### Visualizing regularization

#

# Below is a visualization of what happens when you apply regularization. The general idea is that you are \*\*restricting the allowed values of your coefficients\*\* to a certain "region". \*\*Within that region\*\*, you want to find the coefficients that result in the best model.

# ![Lasso and Ridge Coefficient Plots](images/lasso\_ridge\_coefficients.png)

# In this diagram:

#

# - We are fitting a linear regression model with \*\*two features\*\*, $x\_1$ and $x\_2$.

# - $\hat\beta$ represents the set of two coefficients, $\beta\_1$ and $\beta\_2$, which minimize the RSS for the \*\*unregularized model\*\*.

# - Regularization restricts the allowed positions of $\hat\beta$ to the \*\*blue constraint region:\*\*

# - For lasso, this region is a \*\*diamond\*\* because it constrains the absolute value of the coefficients.

# - For ridge, this region is a \*\*circle\*\* because it constrains the square of the coefficients.

# - The \*\*size of the blue region\*\* is determined by $\alpha$, with a smaller $\alpha$ resulting in a larger region:

# - When $\alpha$ is zero, the blue region is infinitely large, and thus the coefficient sizes are not constrained.

# - When $\alpha$ increases, the blue region gets smaller and smaller.

#

# In this case, $\hat\beta$ is \*\*not\*\* within the blue constraint region. Thus, we need to \*\*move $\hat\beta$ until it intersects the blue region\*\*, while \*\*increasing the RSS as little as possible.\*\*

#

# From page 222 of [An Introduction to Statistical Learning](http://www-bcf.usc.edu/~gareth/ISL/):

#

# > The ellipses that are centered around $\hat\beta$ represent \*\*regions of constant RSS\*\*. In other words, all of the points on a given ellipse share a common value of the RSS. As the ellipses expand away from the least squares coefficient estimates, the RSS increases. Equations (6.8) and (6.9) indicate that the lasso and ridge regression coefficient estimates are given by the \*\*first point at which an ellipse contacts the constraint region\*\*.

#

# > Since \*\*ridge regression\*\* has a circular constraint with no sharp points, this intersection will not generally occur on an axis, and so the ridge regression coefficient estimates will be exclusively non-zero. However, the \*\*lasso\*\* constraint has corners at each of the axes, and so the ellipse will often intersect the constraint region at an axis. When this occurs, one of the coefficients will equal zero. In higher dimensions, many of the coefficient estimates may equal zero simultaneously. In Figure 6.7, the intersection occurs at $\beta\_1 = 0$, and so the resulting model will only include $\beta\_2$.

# ## Part 4: Regularized regression in scikit-learn

# - Communities and Crime dataset from the UCI Machine Learning Repository: [data](http://archive.ics.uci.edu/ml/machine-learning-databases/communities/communities.data), [data dictionary](http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime)

# - \*\*Goal:\*\* Predict the violent crime rate for a community given socioeconomic and law enforcement data