Tutorial: Lasso Regression

Cheat Sheet and Practice Problems

ES335 - Machine Learning
IIT Gandhinagar

July 23, 2025

1 Summary from Slides

1.1 Introduction to Lasso

What is Lasso?

- LASSO: **Least Absolute Shrinkage and Selection Operator**
- Popular regularization technique for linear regression
- Leads to sparse solutions (automatic feature selection)
- Uses L1 regularization penalty

1.2 Objective Function

Constrained Form:

$$\boldsymbol{\theta}_{\text{opt}} = \arg\min_{\boldsymbol{\theta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) \text{ subject to } ||\boldsymbol{\theta}||_1 < s$$

Unconstrained Form (using KKT conditions):

$$\boldsymbol{\theta}_{\text{opt}} = \arg\min_{\boldsymbol{\theta}} \{(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \delta^2 ||\boldsymbol{\theta}||_1\}$$

where δ^2 is the regularization parameter and $||\boldsymbol{\theta}||_1 = \sum_{j=0}^{d} |\theta_j|$

1.3 Key Properties

Sparsity: Lasso sets coefficients of less important features to exactly zero
Feature Selection: Automatically performs variable selection
Convex but Non-differentiable: The L1 penalty is not differentiable at zero

1.4 Solving Lasso: Coordinate Descent

Since the L1 penalty is not differentiable everywhere, we cannot use standard gradient descent. Instead, we use **coordinate descent**:
Key Idea:

- Optimize one parameter at a time while keeping others fixed
- Turns multi-dimensional problem into series of one-dimensional problems
- No step-size selection needed
- Converges for Lasso objective

1.5 Coordinate Descent Algorithm

For each parameter θ_j:

Step 1: Compute ρ_j and z_j:

$$\rho_j = \sum_{i=1}^{n} x_i^j (y_i - \hat{y}_i^{(-j)})$$

$$z_j = \sum_{i=1}^{n} (x_i^j)^2$$

where $\hat{y}_i^{(-j)}$ is the prediction without the j-th parameter.

Step 2: Apply soft-thresholding:

$$\theta_j = \begin{cases} \frac{\rho_j + \frac{\delta^2}{2}}{z_j} & \text{if } \rho_j < -\frac{\delta^2}{2} \\ 0 & \text{if } -\frac{\delta^2}{2} \leq \rho_j \leq \frac{\delta^2}{2} \\ \frac{\rho_j - \frac{\delta^2}{2}}{z_j} & \text{if } \rho_j > \frac{\delta^2}{2} \end{cases}$$

1.6 Subgradients

For non-differentiable functions, we use **subgradients**:

For $f(x) = |x|$:

$$\frac{\partial |x|}{\partial x} = \begin{cases} 1 & \text{if } x > 0 \\ [-1, 1] & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}$$

1.7 Lasso vs Ridge Comparison

Property	**Ridge (L2)**	**Lasso (L1)**
Penalty	$\sum_j \theta_j^2$	$\sum_j \lvert\theta_j\rvert$
Sparsity	No	Yes
Feature Selection	No	Yes
Differentiable	Yes	No
Solution Method	Closed form	Coordinate descent

2 Practice Problems

Problem : Basic Understanding

Explain the difference between Ridge and Lasso regression in terms of: a) The penalty term used b) The effect on parameter values c) Feature selection capability

Problem : Objective Function

Write the complete Lasso objective function for a dataset with n samples and d features. Define all terms clearly.

Problem : L1 Norm Calculation

Given $\boldsymbol{\theta} = [2, -3, 0, 1.5, -0.5]$, calculate: a) The L1 norm $||\boldsymbol{\theta}||_1$ b) The L2 norm $||\boldsymbol{\theta}||_2$ c) How many non-zero parameters are there?

Problem : Subgradient Calculation

For the function $f(x) = |x|$, find the subgradient at: a) $x = 3$ b) $x = -2$ c) $x = 0$
Explain why the subgradient at $x = 0$ is an interval rather than a single value.

Problem : Coordinate Descent Setup

For a simple dataset with one feature:

$$X = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 1 \end{bmatrix}, \quad Y = \begin{bmatrix} 4 \\ 5 \\ 3 \end{bmatrix}$$

Set up the coordinate descent algorithm. Calculate ρ_0, ρ_1, z_0, and z_1 for the initial iteration with $\theta_0 = 0$, $\theta_1 = 0$.

Problem : Soft Thresholding

Apply the soft-thresholding operator with $\delta^2 = 2$ to the following ρ values: a) $\rho = 3$ with $z = 4$ b) $\rho = -0.5$ with $z = 2$ c) $\rho = 1.5$ with $z = 3$
Calculate the resulting θ values using the Lasso coordinate descent update rule.

Problem : Sparsity Analysis

Explain why Lasso regression produces sparse solutions while Ridge regression does not. Use the geometric interpretation involving the constraint regions.

Problem : Regularization Parameter Selection

You have a Lasso model with varying regularization parameter δ^2: - $\delta^2 = 0$: All features retained, MSE = 0.1 - $\delta^2 = 1$: 8/10 features retained, MSE = 0.15 - $\delta^2 = 5$: 3/10 features retained, MSE = 0.25 - $\delta^2 = 10$: 1/10 features retained, MSE = 0.45
Which value would you choose and why? Consider the bias-variance tradeoff.

Problem : Manual Coordinate Descent

Perform one complete iteration of coordinate descent for the Lasso problem:
Dataset: $(x_1, y_1) = (1, 2)$, $(x_2, y_2) = (2, 3)$
Model: $y = \theta_1 x$ (no intercept)
Initial: $\theta_1 = 1$, $\delta^2 = 0.5$
Calculate the updated θ_1 value.

Problem : Feature Selection Scenario

You have a dataset with 1000 features but only 100 samples. Explain: a) Why this is challenging for ordinary least squares b) How Lasso regression can help c) What you expect to happen to the parameter estimates

Problem : Convergence Analysis

In coordinate descent for Lasso: a) Why don't we need to choose a step size? b) What determines the order of coordinate updates? c) How do we know when the algorithm has converged?

Problem : Regularization Path

Describe what happens to the Lasso solution as δ^2 increases from 0 to infinity: a) At $\delta^2 = 0$ b) For small positive δ^2 c) For large δ^2 d) At $\delta^2 \to \infty$
Sketch the regularization path showing how parameters change.

Problem : Computational Comparison

Compare the computational complexity of solving: a) Ordinary least squares: $(X^T X)^{-1} X^T Y$ b) Ridge regression: $(X^T X + \delta^2 I)^{-1} X^T Y$ c) Lasso regression using coordinate descent
When is each method preferred?

Problem : Multi-dimensional Soft Thresholding

For a 3-feature problem with current values:

$$\rho = [2.5, -0.8, 1.2], \quad z = [4, 2, 3], \quad \delta^2 = 1.5$$

Apply coordinate descent to update all three parameters θ_0, θ_1, θ_2.

Problem : Real-world Application

You're building a model to predict house prices with 50 potential features (size, location, age, etc.). Some features may be irrelevant or redundant.
a) Explain why Lasso might be preferred over ordinary least squares b) How would you choose the regularization parameter? c) How would you interpret the final model with only 8 non-zero coefficients?