Precept 5: Regularization in linear models and Hyperparameter tuning using cross-validation

COS424/524/SML302 Spring 2021

Xiaoyan Li, Yaniv Ovadia

Topics:

- Regularization in linear regression and logistic regression
  - (1) Ridge (L2 penalty)
  - (2) Lasso (L1 penalty)
  - (3) Elastic Net (A combination of L1 and L2 penalty)
- Hyperparameter tuning using cross-validation
- Missing data
- Paper for COS524

Ridge Regression

- Minimization objective:
  - Sum of squared residuals + ***α*** *(sum of square of coefficients):

$$||Y - X\beta||_2^2 + \alpha||\beta||_2^2$$

  - L2 regularization:
    - ***α*** is a regularization/shrinkage parameter, controls the size of the coefficient, the strength of regularization

$$\arg\max_{\beta} P(\beta|D) = \arg\max_{\beta} \log \overset{\text{posterior}}{P(\beta|D)}$$

$$= \arg\max_{\beta} \log \overset{\text{likelihood}}{P(D|\beta)} + \log \overset{\text{prior}}{P(\beta)} - \overset{\text{irrelevant}}{\cancel{\log P(D)}}$$

$$= \arg\max_{\beta} \log \left[\prod_{i}^{n} N(y_i | x_i^T\beta, \sigma^2) \right] + \log N(\beta | 0, \tau^2)$$

$$= \arg\max_{\beta} \log \left[\prod_{i}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ \frac{-1}{2\sigma^2} (y_i - x_i^T\beta)^2 \right\} \right] + \log N(\beta | 0, \tau^2)$$

$$= \arg\max_{\beta} n \underset{\text{irrelevant}}{\cancel{\log \frac{1}{\sigma\sqrt{2\pi}}}} + \sum_{i}^{n} \log \exp\left\{ - \text{''} - \right\} + \log N(\beta | 0, \tau^2)$$

$$= \arg\max_{\beta} \sum_{i}^{n} \frac{-1}{2\sigma^2} (y_i - x_i^T\beta)^2 + \log \prod_{j}^{m} \underset{\text{irrelevant}}{\cancel{\frac{1}{\tau\sqrt{2\pi}}}} \exp\left\{ \frac{-1}{2\tau^2} \beta_j^2 \right\}$$

$$= \arg\max_{\beta} \quad \text{———} \, \text{''} \, \text{———} \quad + \sum_{j}^{m} \left(\frac{-1}{2\tau^2} \right) \beta_j^2$$

$$= \arg\max_{\beta} \quad \text{———} \, \text{''} \, \text{———} \quad + \left(\frac{-1}{2\tau^2} \right) \|\beta\|_2^2$$

$$= \arg\underline{\min}_{\beta} \frac{1}{2\sigma^2} \sum_{i}^{n} (y_i - x_i^T\beta)^2 + \frac{1}{2\tau^2} \|\beta\|_2^2$$

$$= \arg\underline{\min}_{\beta} \sum_{i}^{n} (y_i - x^T\beta)^2 + \underbrace{\frac{\sigma^2}{\tau^2}}_{=\lambda_2} \|\beta\|_2^2$$

posterior

$$\arg\max_{\beta} P(\beta|D) = \arg\max_{\beta} \log P(\beta|D)$$

likelihood prior

irrelevant

$$= \arg\max_{\beta} \log P(D|\beta) + \log P(\beta) - \cancel{\log P(D)}$$

$$= \arg\max_{\beta} \log \left[\prod_{i}^{n} N(y_i | x_i^T\beta, \sigma^2) \right] + \log N(\beta | 0, \tau^2)$$

$$= \arg\max_{\beta} \log \left[\prod_{i}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ \frac{-1}{2\sigma^2} (y_i - x_i^T\beta)^2 \right\} \right] + \log N(\beta | 0, \tau^2)$$

$$= \arg\max_{\beta} n \log \cancel{\frac{1}{\sigma\sqrt{2\pi}}} + \sum_{i}^{n} \log \exp\left\{ -\text{"}- \right\} + \log N(\beta | 0, \tau^2)$$

irrelevant

$$= \arg\max_{\beta} \sum_{i}^{n} \frac{-1}{2\sigma^2} (y_i - x_i^T\beta)^2 + \log \prod_{j}^{m} \cancel{\frac{1}{\tau\sqrt{2\pi}}} \exp\left\{ \frac{-1}{2\tau^2} \beta_j^2 \right\}$$

irrelevant

$$= \underset{\beta}{\arg\max}\; n \log \underbrace{\cancel{\frac{1}{\sigma\sqrt{2\pi}}}}_{\text{irrelevant}} + \sum_{i}^{n} \log \exp\left\{ -\text{''}- \right\} + \log N(\beta \mid 0, \tau^2)$$

$$= \underset{\beta}{\arg\max} \sum_{i}^{n} \frac{-1}{2\sigma^2}\left(y_i - x_i^T\beta\right)^2 + \log \prod_{j}^{m} \underbrace{\cancel{\frac{1}{\tau\sqrt{2\pi}}}}_{\text{irrelevant}} \exp\left\{ \frac{-1}{2\tau^2}\beta_j^2 \right\}$$

$$= \underset{\beta}{\arg\max} \quad \text{———}\ \text{''}\ \text{———} + \sum_{j}^{m} \left(\frac{-1}{2\tau^2}\right)\beta_j^2$$

$$= \underset{\beta}{\arg\max} \quad \text{———}\ \text{''}\ \text{———} + \left(\frac{-1}{2\tau^2}\right)\|\beta\|_2^2$$

$$= \underset{\beta}{\arg\min}\; \frac{1}{2\sigma^2}\sum_{i}^{n}\left(y_i - x_i^T\beta\right)^2 + \frac{1}{2\tau^2}\|\beta\|_2^2$$

$$= \underset{\beta}{\arg\min} \sum_{i}^{n}\left(y_i - x_i^T\beta\right)^2 + \underbrace{\frac{\sigma^2}{\tau^2}}_{=\lambda_2}\|\beta\|_2^2$$

Ridge regression for 10 different values of α ranging from 1e-15 to 20.

	rss	intercept	coef_x_1	coef_x_2	coef_x_3	coef_x_4	coef_x_5	coef_x_6	coef_x_7	coef_x_8	coef_x_9	coef_x_10	coef_x_11	co
alpha_1e-15	0.87	95	-3e+02	3.8e+02	-2.4e+02	66	0.96	-4.8	0.64	0.15	-0.026	-0.0054	0.00086	0.0
alpha_1e-10	0.92	11	-29	31	-15	2.9	0.17	-0.091	-0.011	0.002	0.00064	2.4e-05	-2e-05	-4.
alpha_1e-08	0.95	1.3	-1.5	1.7	-0.68	0.039	0.016	0.00016	-0.00036	-5.4e-05	-2.9e-07	1.1e-06	1.9e-07	2e-
alpha_0.0001	0.96	0.56	0.55	-0.13	-0.026	-0.0028	-0.00011	4.1e-05	1.5e-05	3.7e-06	7.4e-07	1.3e-07	1.9e-08	1.9
alpha_0.001	1	0.82	0.31	-0.087	-0.02	-0.0028	-0.00022	1.8e-06	1.2e-06	3.4e-06	7.3e-07	1.3e-07	1.9e-08	1.7
alpha_0.01	1.4	1.3	-0.088	-0.052	-0.01	-0.0014	-0.00013	7.2e-07	4.1e-06	1.3e-06	3e-07	5.6e-08	9e-09	1.1
alpha_1	5.6	0.97	-0.14	-0.019	-0.003	-0.00047	-7e-05	-9.9e-06	-1.3e-06	-1.4e-07	-9.3e-09	1.3e-09	7.8e-10	2.4
alpha_5	14	0.55	-0.059	-0.0085	-0.0014	-0.00024	-4.1e-05	-6.9e-06	-1.1e-06	-1.9e-07	-3.1e-08	-5.1e-09	-8.2e-10	-1.
alpha_10	18	0.4	-0.037	-0.0055	-0.00095	-0.00017	-3e-05	-5.2e-06	-9.2e-07	-1.6e-07	-2.9e-08	-5.1e-09	-9.1e-10	-1.
alpha_20	23	0.28	-0.022	-0.0034	-0.0006	-0.00011	-2e-05	-3.6e-06	-6.6e-07	-1.2e-07	-2.2e-08	-4e-09	-7.5e-10	-1.

Linear Regression:

model_pow_15	0.7	-3.6e+04	2.4e+05	-7.5e+05	1.4e+06	-1.7e+06	1.5e+06	-1e+06	5e+05	-1.9e+05	5.4e+04	-1.2e+04	1.9e+03	[illegible]

Q: What observations do you have based on this table?

Ridge regression for 10 different values of α ranging from 1e-15 to 20.

	rss	intercept	coef_x_1	coef_x_2	coef_x_3	coef_x_4	coef_x_5	coef_x_6	coef_x_7	coef_x_8	coef_x_9	coef_x_10	coef_x_11	co
alpha_1e-15	0.87	95	-3e+02	3.8e+02	-2.4e+02	66	0.96	-4.8	0.64	0.15	-0.026	-0.0054	0.00086	0.0
alpha_1e-10	0.92	11	-29	31	-15	2.9	0.17	-0.091	-0.011	0.002	0.00064	2.4e-05	-2e-05	-4.
alpha_1e-08	0.95	1.3	-1.5	1.7	-0.68	0.039	0.016	0.00016	-0.00036	-5.4e-05	-2.9e-07	1.1e-06	1.9e-07	2e-
alpha_0.0001	0.96	0.56	0.55	-0.13	-0.026	-0.0028	-0.00011	4.1e-05	1.5e-05	3.7e-06	7.4e-07	1.3e-07	1.9e-08	1.9
alpha_0.001	1	0.82	0.31	-0.087	-0.02	-0.0028	-0.00022	1.8e-05	1.2e-05	3.4e-06	7.3e-07	1.3e-07	1.9e-08	1.7
alpha_0.01	1.4	1.3	-0.088	-0.052	-0.01	-0.0014	-0.00013	7.2e-07	4.1e-06	1.3e-06	3e-07	5.6e-08	9e-09	1.1
alpha_1	5.6	0.97	-0.14	-0.019	-0.003	-0.00047	-7e-05	-9.9e-06	-1.3e-06	-1.4e-07	-9.3e-09	1.3e-09	7.8e-10	2.4
alpha_5	14	0.55	-0.059	-0.0085	-0.0014	-0.00024	-4.1e-05	-6.9e-06	-1.1e-06	-1.9e-07	-3.1e-08	-5.1e-09	-8.2e-10	-1.
alpha_10	18	0.4	-0.037	-0.0055	-0.00095	-0.00017	-3e-05	-5.2e-06	-9.2e-07	-1.6e-07	-2.9e-08	-5.1e-09	-9.1e-10	-1.
alpha_20	23	0.28	-0.022	-0.0034	-0.0006	-0.00011	-2e-05	-3.6e-06	-6.6e-07	-1.2e-07	-2.2e-08	-4e-09	-7.5e-10	-1.

model_pow_15	0.7	-3.6e+04	2.4e+05	-7.5e+05	1.4e+06	-1.7e+06	1.5e+06	-1e+06	5e+05	-1.9e+05	5.4e+04	-1.2e+04	1.9e+03

A: observations:

- RSS increases as alpha increases;
- High alpha values can lead to underfitting
- Even the smallest alpha gives us significant reduction in magnitude of coefficients;
- Though the coefficients are very small, they are NOT zero.

Topics:

- Regularization in linear regression and logistic regression
  - (1) Ridge (L2 penalty)
  - (2) Lasso (L1 penalty)
  - (3) Elastic Net (A combination of L1 and L2 penalty)
- Hyperparameter tuning using cross-validation
- Missing data
- Paper for COS524

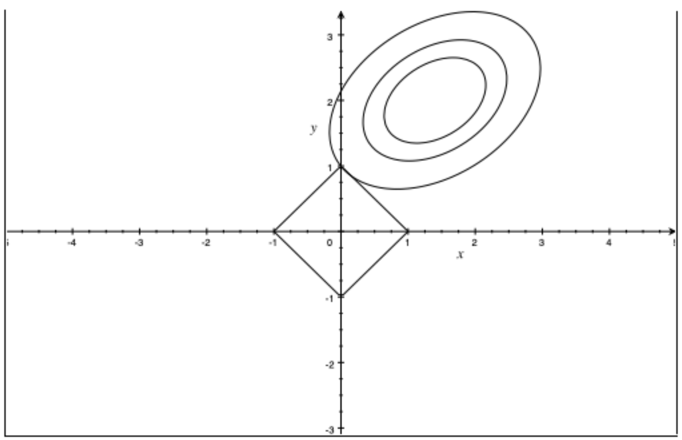
Lasso Regression

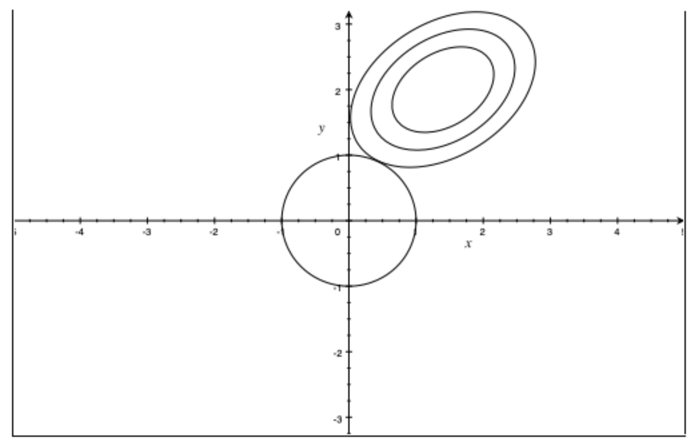
- Minimization objective:
  - Sum of squared residuals + $\boldsymbol{\alpha}$ *(sum of absolute value of coefficients):

$$||Y - X\beta||_2^2 + \alpha||\beta||_1$$

  - L1 regularization:
    - $\boldsymbol{\alpha}$ is a regularization/shrinkage parameter, controls the **sparsity** of the parameters, the strength of regularization
    - Can be derived as a MAP estimate with Laplace distribution prior

Intuition for Sparsity of L1-Regularization





Lasso regression for 10 different values of α ranging from 1e-15 to 10.

	rss	intercept	coef_x_1	coef_x_2	coef_x_3	coef_x_4	coef_x_5	coef_x_6	coef_x_7	coef_x_8	coef_x_9	coef_x_10	coef_x_11	coe
alpha_1e-15	0.96	0.22	1.1	-0.37	0.00089	0.0016	-0.00012	-6.4e-05	-6.3e-06	1.4e-06	7.8e-07	2.1e-07	4e-08	5.4
alpha_1e-10	0.96	0.22	1.1	-0.37	0.00088	0.0016	-0.00012	-6.4e-05	-6.3e-06	1.4e-06	7.8e-07	2.1e-07	4e-08	5.4
alpha_1e-08	0.96	0.22	1.1	-0.37	0.00077	0.0016	-0.00011	-6.4e-05	-6.3e-06	1.4e-06	7.8e-07	2.1e-07	4e-08	5.3
alpha_1e-05	0.96	0.5	0.6	-0.13	-0.038	-0	0	0	0	7.7e-06	1e-06	7.7e-08	0	0
alpha_0.0001	1	0.9	0.17	-0	-0.048	-0	-0	0	0	9.5e-06	5.1e-07	0	0	0
alpha_0.001	1.7	1.3	-0	-0.13	-0	-0	-0	0	0	0	0	0	1.5e-08	7.5
alpha_0.01	3.6	1.8	-0.55	-0.00056	-0	-0	-0	-0	-0	-0	-0	0	0	0
alpha_1	37	0.038	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0
alpha_5	37	0.038	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0
alpha_10	37	0.038	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0

HIGH SPARSITY

Q: What observations do you have based on this table?

Lasso regression for 10 different values of α ranging from 1e-15 to 10.

	rss	intercept	coef_x_1	coef_x_2	coef_x_3	coef_x_4	coef_x_5	coef_x_6	coef_x_7	coef_x_8	coef_x_9	coef_x_10	coef_x_11	coe
alpha_1e-15	0.96	0.22	1.1	-0.37	0.00089	0.0016	-0.00012	-6.4e-05	-6.3e-06	1.4e-06	7.8e-07	2.1e-07	4e-08	5.4
alpha_1e-10	0.96	0.22	1.1	-0.37	0.00088	0.0016	-0.00012	-6.4e-05	-6.3e-06	1.4e-06	7.8e-07	2.1e-07	4e-08	5.4
alpha_1e-08	0.96	0.22	1.1	-0.37	0.00077	0.0016	-0.00011	-6.4e-05	-6.3e-06	1.4e-06	7.8e-07	2.1e-07	4e-08	5.3
alpha_1e-05	0.96	0.5	0.6	-0.13	-0.038	-0	0	0	0	7.7e-06	1e-06	7.7e-08	0	0
alpha_0.0001	1	0.9	0.17	-0	-0.048	-0	-0	0	0	9.5e-06	5.1e-07	0	0	0
alpha_0.001	1.7	1.3	-0	-0.13	-0	-0	-0	0	0	0	0	0	1.5e-08	7.5
alpha_0.01	3.6	1.8	-0.55	-0.00056	-0	-0	-0	-0	-0	-0	-0	0	0	0
alpha_1	37	0.038	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0
alpha_5	37	0.038	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0
alpha_10	37	0.038	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0

HIGH SPARSITY

A: observations:

- RSS increases as alpha increases;
- Many of the coefficients are zero even for vey small values of alpha.

Limitations of Lasso Regression

1. Number of samples (N) < number of features (P):
   - Lasso selects at most N features.
2. Highly correlated features
   - Lasso tends to select one and ignore the others
   - Not necessarily consistent between fits

Topics:

- Regularization in linear regression and logistic regression
  - (1) Ridge (L2 penalty)
  - (2) Lasso (L1 penalty)
  - (3) Elastic Net (A combination of L1 and L2 penalty)
- Hyperparameter tuning using cross-validation
- Missing data
- Paper for COS524

Elastic Net Regression

- Objective:
  - NLL $+ \lambda_1 ||\beta||_1 + \lambda_2 ||\beta||_2^2$
  - A combination of L1 and L2 regularization:
    - Add 2 penalty terms:
    - $\boldsymbol{\lambda}$1 and $\boldsymbol{\lambda}_2$ are regularization/shrinkage parameters, controls the strength of regularization
    - pre-chosen hyperparameters for the model
    - Ridge and Lasso regressions are special cases of it.
    - $\boldsymbol{\lambda}$1 and $\boldsymbol{\lambda}$2 can be controled together or separtely.
      - Eg. *alpha=1.0, l1_ratio=0.5 in* sklearn.linear_model.ElasticNet

Regularized Logistic Regression

- Can perform regularization on Logistic regression, similar to regularized linear regression (They are all in the generalized linear model framework; Logistic regression is a classifier.)
- Add penalty term to object function
  - L1-penalty
  - L2-penalty
  - Elastic net penalty

Topics:

- Regularization in linear regression and logistic regression
  - (1) Ridge (L2 penalty)
  - (2) Lasso (L1 penalty)
  - (3) Elastic Net (A combination of L1 and L2 penalty)
- Hyperparameter tuning using cross-validation
- Missing data
- Paper for COS524

What are hyperparameters?

- Pre-selected values for some model parameters before fitting the model on the training data
- Examples:
  - $\boldsymbol{\alpha}$ in Ridge and Lasso Regression
  - $\boldsymbol{\lambda}1$ and $\boldsymbol{\lambda}_2$ in Elastic Net Regression
  - K in K-nearest neighbors classifier
  - Kernel parameter and penalty parameter in SVM
  - …

Hyperparameter Tuning using Cross-Validation

- (1) Set K=k for K-fold cross-validation
- (2) Random split training data set D into k folds: S_1, S_2, … , S_k
- (3) For C in {C_1 C_2, … , C_m}, assume C is the hyperparameter
  - For i = 1, 2, … , k
    - Let fold i (S_i) be the test(held out) fold.
    - Fit the model on the other k-1 folds.
    - Predict on the test fold S_i.
  - Compute generalization error (E_c) from one prediction for each sample
- (4) $C^* = \underset{c}{\mathrm{argmin}} Ec$
- (5) retrain your model on the training data set D with C*

- Q1: Can step 2 be moved inside the for-loop?
- Q2: How do you choose {C_1 C_2, … , C_m}?
- Q3: Can you use the test data to set your hyperparameters?
- Q4: Can we run either of these loops in parallel?

Hyperparameter Tuning using Cross-Validation

- (1) Set K=k for K-fold cross-validation
- (2) Random split training data set D into k folds: $S_1, S_2, \ldots, S_k$
- (3) For C in $\{C_1\ C_2, \ldots, C_m\}$, assume C is the hyperparameter
  - For i = 1, 2, … , k
    - Let fold i (S_i) be the test(held out) fold.
    - Fit the model on the other k-1 folds.
    - Predict on the test fold S_i.
  - Compute generalization error (E_c) from one prediction for each sample
- (4) $C^* = \underset{c}{\operatorname{argmin}} Ec$
- Q1: Can step 2 be moved inside the for-loop?
  - A: No. should keep the same folds.
- Q2: how do you choose $\{C_1\ C_2, \ldots, C_m\}$?
  - A: grid search, or random selection in the parameter space, others…
- Q3: Can you use the leader board to set your hyperparameters?
  - NO
- Q4: Can we run either of these loops in parallel?
  - A: Yes, both!

Some Classes in sklearn for regularization and hyperparameter tuning

- sklearn.linear_model.Ridge
- sklearn.linear_model.RidgeCV
- sklearn.linear_model.Lasso
- sklearn.linear_model.LassoCV
- sklearn.linear_model.ElasticNet
- sklearn.linear_model.ElasticNetCV
- sklearn.linear_model.SGDClassifier
- sklearn.linear_model.SGDRegressor
- sklearn.linear_model.LogisticRegression
- sklearn.linear_model.LogisticRegressionCV
  - SGD – training with stochastic gradient decent
  - CV– with built-in cross-validation

Paper:

- Group discussion for COS524 in breakout rooms (~15 minutes)
- Share your opinions in the shared google docs:
  - Will send you the links in chat
  - Go to the link for your breakout room
  - You can focus on one of the questions
- Come back for precept wrap up

Paper:

- 1.
- 2
- 3

Wrap up for Precept 3

regularization in linear regression and logistic regression

- (1) Ridge (L2 penalty)
- (2) Lasso (L1 penalty)
- (3) Elastic Net (A combination of L1 and L2 penalty)

- Missing data
- Paper for COS524

Resources:(many graphs and texts are taken from the following resources.)

- “A Complete Tutorial on Ridge and Lasso Regression in Python” by Aarshay Jain https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-ridge-lasso-regression-python/

- "Missing data" by Iris Eekhout
  - https://www.iriseekhout.com/missing-data/