Hyperparameter Optimization and Bilevel Programming

Wei Jiang
Chia-Hsiu Chang (Todd)

Date: Dec. 08. 2016

Outline

- Hyperparameter Optimization
- Support Vector Machine Classification
- Logistic Regression

Hyperparameter Optimization

- Machine learning：

  Contain a set of hyperparameters for controlling the model complexity to prevent overfitting.

- Model selection：

  Choosing the best hyperparameters to maximize the model performance on unseen test data.

- Cross validation：

  Evaluate how well model will generalize to an unseen test data.

Methods

- Bilevel optimization (software: GAMS)

  Leader：Minimize the out-of-sample error

  Follower：Optimize the in-sample-error for each fold

- Grid search

  Traditional way, exhaustive searching.

Support Vector Machine

Primal:

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_n$$

$$\text{s.t.}\ \ y_i\left(x_i^T\beta + \beta_0\right) \geq 1 - \xi_i, \qquad \forall i$$

$$\xi_i \geq 0$$

Dual:

$$\min_{\alpha_i} \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i^T x_j$$

$$\text{s.t.}\quad 0 \leq \alpha_i \leq C, \quad \forall i$$

$$\textstyle\sum_{i=1}^{N}\alpha_i y_i = 0$$

I solved the dual problem in GAMS as a QCP using CPLEX

Why dual: Straightforward to apply kernel tricks. A nicer quadratic program.

Support Vector Machine

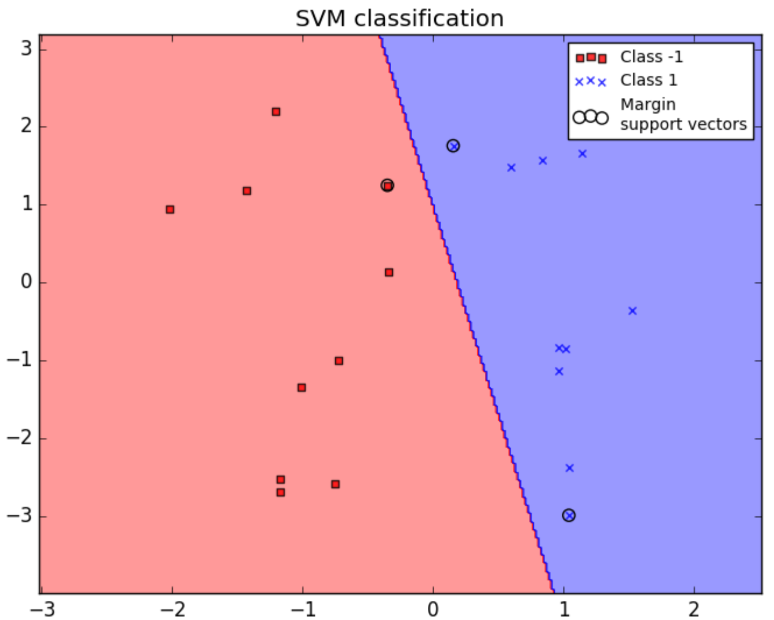


Figure 1.Randomly generated binary class data. Sample size: 20, number of features: 2. Two classes are linear separable. So SVM yields hard margin.

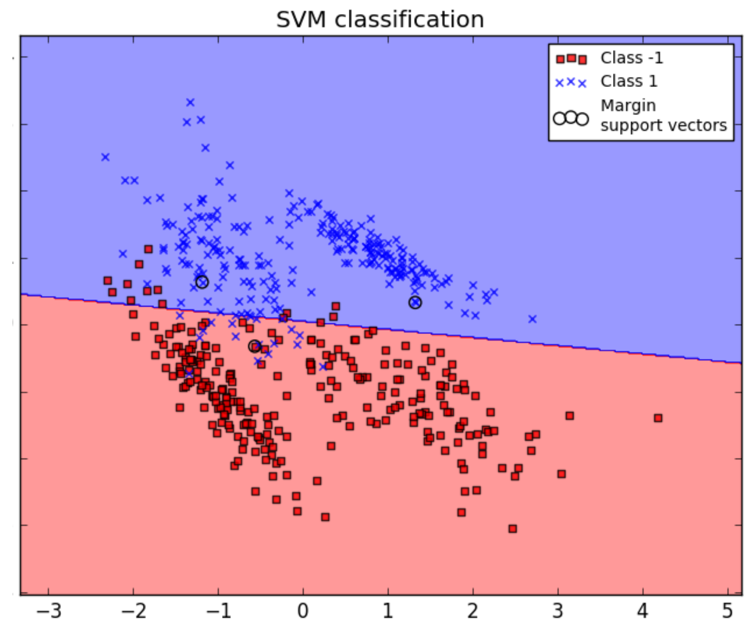


Figure 2. 1.Randomly generated binary class data. Sample size: 500, number of features: 2. Two classes are linear non-separable. So SVM yields soft margin, which allows classification errors.

All my code (SVM, SVM bilevel) publicly available on github: https://github.com/wjiang16/SVM_MPEC
(Code in python, called GAMS as the optimization solver through GAMS python API)

Upper Level Problem

(Minimize cross-validation misclassification error)

$$\min \theta(\beta, \beta_0, C) = \frac{1}{T}\sum_{t=1}^{T} \frac{1}{|N_t|} \sum_{i \in N_t} [-y_i(x_i^T\beta + \beta_0)]^*$$

A step function:

$$(b^*)_i = \begin{cases} 1, & b_i > 0 \\ 0, & b_i \leq 0 \end{cases}$$

b_i^* is the solution of the following linear program:

$$b_i^* = argmin_{b_i} \;\; b_i y_i (x_i^T\beta + \beta_0)$$
$$0 \leq b_i \leq 1\,, \quad \forall i \in N_t$$

Bilevel Problem

$$\min_{C, b_i, \beta^t, \beta_0^t, \lambda_i^t, \alpha_i^t, \xi_i^t} \quad \frac{1}{T} \sum_{t=1}^{T} \frac{1}{|N_t|} \sum_{i \in N_t} b_i^t$$

$\text{s.t.} \quad \text{for t} = 1, \ldots \ldots \text{T}$

$$\begin{array}{c} 0 \le b_i^t \perp y_i(x_i^T \beta + \beta_0) + \lambda_i^t \ge 0 \\ 0 \le \lambda_i^t \perp 1 - b_i^t \ge 0 \end{array} \quad \forall i \in N_t$$

Missclassification error

$$\begin{array}{c} 0 \le C - \alpha_i \perp \xi_i \ge 0 \\ 0 \le \alpha_i \perp \left[y_i \left(x_i^T \beta + \beta_0 \right) - (1 - \xi_i) \right] \ge 0 \\ \beta^t = \sum_{i \in \overline{N_t}} \alpha_i y_i x_i \\ \sum_{i \in \overline{N_t}} \alpha_i y_i = 0 \end{array} \quad \forall i \in \overline{N_t}$$

SVM

Primal Dual Formulation
(using Slater's condition)

$$\min_{C,b_i,\beta^t,\beta_0^t,\lambda_i^t,\alpha_i^t,\xi_i^t} \frac{1}{T}\sum_{t=1}^{T}\frac{1}{|N_t|}\sum_{i\in N_t} b_i^t$$

$$\begin{aligned}
\text{strong duality : }\ & b_i y_i(x_i^T\beta+\beta_0) = -\lambda_i \\
& 0 \le b_i \le 1 \qquad\qquad \forall i \in N_t \\
& -\lambda_i \le y_i\left(x_i^T\beta+\beta_0\right) \\
& \lambda_i \ge 0
\end{aligned}$$

$$\begin{aligned}
\text{strong duality : }\ & \textstyle\sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_i x_i^T x_{j_i} = \frac{1}{2}\beta^T\beta + C\sum_{i=1}^{N}\xi_i \\
& y_i\left(x_i^T\beta+\beta_0\right) \ge 1-\xi_i \qquad\qquad \forall i \in \overline{N_t} \\
& \xi_i \ge 0 \\
& 0 \le \alpha_i \le C \\
& \textstyle\sum_{i=1}^{N}\alpha_i y_i = 0
\end{aligned}$$

Primal Dual Formulation
(using Slater's condition)

- Equality constraints are difficult to handle for NLP solvers in GAMS, e.g., CONOPT.
- Put equality constraints (duality gap) in objective function, solve this penalization formulation iteratively* to minimize the obj. function using CONOPT.
- Solution will be optimal when all the duality gaps are 0.

Solve this penalization formulation iteratively :* Put a penalty coefficient ϵ_i ($\epsilon_0 = 1$) in front of the duality gaps. For iteration i, set $\epsilon_0 = 10 * \epsilon_{i-1}$, use the solution from iteration $i - 1$ as the initial value for the decision variables.

SVM Classification Results

Table 1. The results of SVM classification models
(Randomly generated data, same data as in Figure 2 on slides 6)

Method	Bilevel optimization (5 folds)	Grid search (5 folds, searches 5 value of C)
*C**	0.1	10
Cross-validation accuracy	93.4%	93.8%
Running time*	54.33 sec	122.42 sec

C:* Grid search results show that for this data set, values of C in the range [0.1, 10] all yield cross-validation accuracy above 93%. Accuracy measure is non-continuous, so there are more than one local optimum.
Running time:* running time in python, including time for data exchange between python and GAMS. Actual GAMS execution time is within 1 second for both cases.

Logistic Regression

$$\log \frac{P\{y = 1|X = x\}}{P\{y = 0|X = x\}} = w^T x$$

- Upper level problem： Minimize the Brier score, a score function that measures the accuracy of probabilistic predictions.

$$(1) \qquad \min \frac{1}{T} \sum_{t=1}^{T} \frac{1}{|N_t|} \sum_{i \in N_t} (P_i - y_i)^2$$

T : Disjoint partitions of data
N_t : Validation sets
P_i : Predicted response
y_i : Real response

- Lower level problem： Maximize the log-Likelihood of Logistic regression.

$$(2) \qquad \max \sum_{j \in \overline{N_t}} \{y_j \ln \mathrm{P}(x_j, w) + (1 - y_j) \ln[1 - \mathrm{P}(x_j, w)]\} - \frac{\lambda}{2} \|w\|^2$$

Equation(2) can be simplified as :

$$(3) \qquad \min \; - \sum_{j \in \overline{N_t}} \{y_j w^T x_j - \ln(1 + e^{-w^T x_j})\} + \frac{\lambda}{2} \|w\|^2$$

$\overline{N_t}$: Training sets
x_i : Observed data
w : Coefficient
λ : Hyperparameter

Logistic Regression

- Take KKT condition of equation (3) and form the bilevel optimization problem：

$$min_{\lambda} \ \frac{1}{T}\sum_{t=1}^{T}\frac{1}{|N_t|}\sum_{i\in N_t}(P_i - y_i)^2$$

$$\text{s.t.}$$

$$\frac{1}{\overline{\overline{N_t}}}\sum_{j\in\overline{N_t}} -x_j\left(y_j - \frac{1}{1+e^{-w^T x_j}}\right) + \lambda\|w\| = 0$$

Data analysis example - logistic regression

- Diabetes Readmission:

  (1) 69973 samples;

  (2) Use 6 observed variables: Gender, Age, Admission, Discharge, Primary diagnosis, Time in hospital;

  (3) Response: 1 = Readmission within 30 days,
  0 = Otherwise.

Data reference：B. Strack *et al*. (2014) Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. *BioMed Research International*, vol. 2014.

Logistic Regression Results

Table 2. The results of logistic regression models

Method	Bilevel optimization		Grid search
	GAMS (2 folds)	GAMS (5 folds)	R (5 folds)
λ	0.0005522	0.0003692	0.0007759
Brier Score	0.0870	0.0870	0.0897
Running time	4.88 sec	6.27 sec	50.12 sec

Reference

- Bennett, K., Kunapuli, G., Hu, J., Pang, J.: Bilevel Optimization and Machine Learning. In: Computational Intelligence: Research Frontiers, no. 5050 in Lecture Notes in Computer Science, pp. 25-47. Springer Berlin Heidelberg (2008)
- Kunapuli, G., Bennett, K., Hu, J., Pang, J.: Bilevel model selection for support vector machines. In: Hansen, P., Pardolos, P. (eds.) CRM Proceedings and Lecture Notes. American Mathematical Society (in press, 2008)
- Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning (2013)

Thank you for listening~

Backup Slides

- We have also tried to present support vector machine regression (SVM regression) problem. We are able to get reasonable results by using a simple regression data (As shown in following slides).

- However, when we use the data of house prices, the results do not make sense.

- We didn’t have enough time to solve this. Hence we present this part in backup slides.

Support Vector Machine Regression

- Upper level problem：Minimize the regularized risk function.

$$(1) \qquad \min \ \frac{1}{T}\sum_{t=1}^{T}\frac{1}{|N_t|}\sum_{i\in N_t}|x_i w^t - y_i|$$

T : Disjoint partitions of data
N_t : Validation sets
x_i : Observed data
y_i : Real response
w : Coefficient

- Lower level problem：Minimize the ε–insentive function.

$$(2) \qquad \arg\min \ \{C\sum_{j\in\overline{N_t}}\max\left(\left|x_j w^t - y_j\right| - \varepsilon, 0\right) + \frac{1}{2}\|w\|^2\}$$

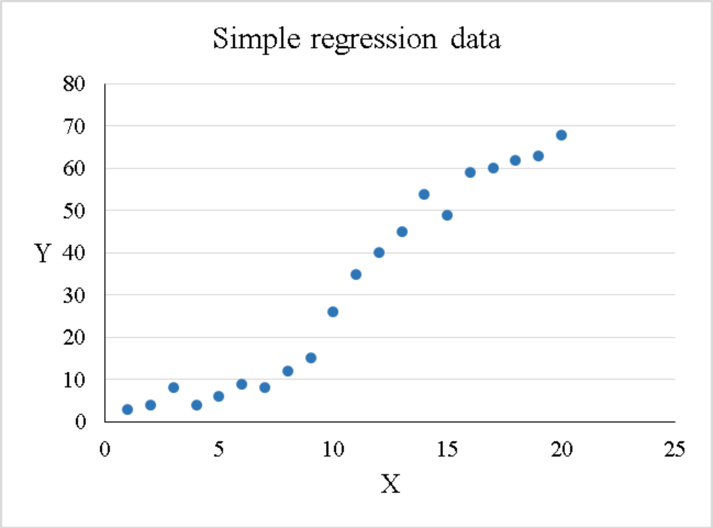
$\overline{N_t}$: Training sets
C, ε : Hyperparameter
C : Regularization parameter
ε : Tube parameter

Reference：Bennett, K. P. *et al*. (2008)
Bilevel Optimization and Machine Learning

SVM Regression

- The tricky part is that the lower level problem contains linear complementarity constraints, so we conduct disjunctive constraints to solve it.

- Data analysis example ：A simple regression example.



Reference: http://www.svm-tutorial.com/2014/10/support-vector-regression-r/

SVM Regression Results

Table 3. The results of simple regression data

Method	Bilevel optimization	Grid search
	GAMS (2 folds)	R (5 folds)
C	3.444	4
ε	0	0.0001
Mean Absolute Deviation	2.303	1.46
Running time	4.24 sec	141.37 sec

SVM Regression

- Data analysis example ：House prices

  (1) 388 samples;

  (2) Observed variables: The total area of the house (ft^2);

  (3) Response: The house prices (USD).

- Since the scale of variables and response are far different, we take square root of the house prices to compute.

Data reference：Cock, D.(2011) Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project. *Journal of Statistics Education*, Vol. 19, No. 3

SVM Regression Results

Table 3. The results of the house prices data

Method	Bilevel optimization
	GAMS (2 folds)
C	0.0000354
ε	161.77
Mean Absolute Deviation	32.122
Running time	4.96 sec