CSC2547: Topics in Statistical Learning Theory Winter 2019

Project Report: Kernel Selection in SVM for binary data

Name: Tian Han Guan *ID: 998978058*

1 Introduction

One of the biggest challenges of using support vector machine (SVM) for non-linear classification problems is how to select an appropriate kernel class. The goal of the project is to compare the generalization bound for the SVM classifier using different kernel class in case where all features are binary. A generalization bound measures a hypothesis class's predictive performance in PAC learning - that is, with high probability, we select function that will have low generalization error.

2 Experiment

The data used is the Mushroom data where the response variable is binary (edible or poisonous), features are all categorical and are encoded to binary variables using one-hot encoding (in total 112 features). Using a train-test split of 70%/30%, the performance of different kernels are shown in Table 1 below. In each case, the optimal hypothesis is obtained using the *Empirical Risk Minimization (ERM) principle* with the *hinge loss function*, where the hyper-parameters γ (which measures the spread of the kernel) is tuned using grid search. The results have shown the following interesting observation:

- Although in practice the Gaussian (Radial Basis Function) kernel has the most popular use, in this case the ranking of the kernels is *Polynomial* > *Sigmoid* > *Gaussian* (based on *generalization error* which represents the estimated population risk of the classifier).

Table 1. Classification performance of various kernel classes.

Kernel Class	Mathematical Definition	Training error	Generalization error	Excess error
Gaussian	$exp(-\gamma\|x-x'\|^2)$	0.04%	8.29%	8.25%
Sigmoid	$tanh(\gamma\langle x,x'\rangle)$	0.39%	3.32%	2.93%
Homogeneous Polynomial (deg = 2)	$(\gamma\langle x,x'\rangle)^2$	0.09%	1.76%	1.67%

3 Theoretical analysis

3.1 Surrogate loss and consistency

First it can be shown that if we can minimize the hinge loss (surrogate for the 0/1 loss), then we can also minimize the 0/1 loss. That is, for all hypothesis in the class,

$$\left|R(\hat{f}) - R(f^*)\right| \leqslant \left|R_l(\hat{f}) - R_l(f^*)\right|$$

In this case, R is the risk w.r.t. the 0/1 loss function and R_l is the risk w.r.t. the hinge loss function. Details can be found in Appendix 1 of the report.

3.2 Rademacher complexity

The next step is to derive an upper bound of Rademacher complexity for each kernel class using Theorem 2 (proved in Appendix 2). Here, let H be a ball with radius r in the RKHS F, $H = \{f \in F : \|f\|_F \leqslant r\}$.

- Gaussian kernel: $R_n(H) \leqslant \frac{r}{\sqrt{n}}$
- Sigmoid kernel: $R_n(H) \leqslant \frac{r\sqrt{tanh(\gamma d)}}{\sqrt{n}}$, d is the dimension of the feature space
- p-degree homogeneous polynomial kernel: $R_n(H) \leqslant \frac{r(\gamma d)^{\frac{p}{2}}}{\sqrt{n}}$

3.3 Generalization bound based on the Rademacher complexity

Using Theorem 3 (proved in Appendix 3), the following generalization bounds are obtained for the 0/1 loss function. With probability $1-\delta$, $\left|R(\hat{f}) - R(f^*)\right| \leqslant \left|R_l(\hat{f}) - R_l(f^*)\right| \leqslant sup_{f\in H}\left|R_l(\hat{f}) - R_l(f)\right| \leqslant M$

- Gaussian kernel: $M = \frac{2r}{\sqrt{n}} + (1+r)\sqrt{\frac{log(\frac{2}{\delta})}{2n}}$
- Sigmoid kernel: $M = \frac{2r\sqrt{tanh(\gamma d)}}{\sqrt{n}} + (1 + r\sqrt{tanh(\gamma d)})\sqrt{\frac{log(\frac{2}{\delta})}{2n}}$
- p-degree homogeneous polynomial kernel: $M = \frac{2r(\gamma d)^{\frac{p}{2}}}{\sqrt{n}} + (1 + r(\gamma d)^{\frac{p}{2}})\sqrt{\frac{log(\frac{2}{\delta})}{2n}}$

3.4 Summary of analysis results

1. Gaussian kernels tend to give higher generalization errors than the polynomial and sigmoid kernels when optimal γ is chosen.
2. When the feature dimension is small relative to the hyper-parameter value (γd is small ($<< 1$)), using a *polynomial kernel* will tend to give lower generalization errors.
3. When the feature dimension is large relative to the hyper-parameter value (γd is large (> 1)), using a *sigmoid kernel* will tend to give lower generalization errors.

In practice it is very difficult to test performance of many kernel classes in fine grid search when data size is large. Using the results above, one can first identify a very rough hyper-parameter space, then choose an appropriate kernel class. This will save much time compared to first obtain the optimal hypothesis from each kernel class and then select the best kernel.

4 Empirical analysis

Table 2 below compares the results of the kernel functions using different hyper-parameter spaces. The following results are observed:

1. The optimal hypothesis of both polynomial and sigmoid kernels have lower generalization errors than the optimal hypothesis of the Gaussian kernel. This is consistent with result 1.
2. When hyper-parameter γ is small as in $(0.01, 0.02)$, (γ is small given the data dimension), polynomial kernel performs better than the sigmoid kernel. However, as γ increases as in $(0.02, 0.05)$ and $(0.05, 0.1)$, sigmoid kernel starts to perform better. This is consistent with result 2 and 3.

Table 2. Generalization error of various kernel functions with different hyper-parameter space ($d = 112$).

Kernel Function	$\gamma \in (0.01, 0.02)$*	$\gamma \in (0.02, 0.05)$	$\gamma \in (0.05, 0.1)$
Gaussian	8.74%	8.99%	8.78%
Sigmoid	5.05%	8.66%	10.26%
Homogeneous Polynomial (deg=2)	1.97%	7.92%	11.24%

*Optimal range of hyper-parameters

Appendix 1. Surrogate loss and consistency

In binary classification problems, the natural choice of loss function is the 0/1 loss, however, the 0/1 loss is non-convex and non-differentiable, therefore makes it very difficult to optimize for the Empirical Risk Minimization (ERM) setting. As a result, a convex surrogate loss function (for example, hinge loss) is usually used as a proxy due to their computational advantages.

Here, one question to ask is how much do we lose by substituting the 0/1 loss by the hinge loss, and can we bound the original generalization risk by the surrogate loss function. The following theorems are being used to show the result.

Theorem 1.1 Let l be a convex loss function. Then l is *classification-calibrated* if and only if

1. l is differentiable at the origin
2. $l'(0) < 0$

It is easy to see that the hinge loss $l(y, f(x)) = max(0, 1 - yf(x))$ is classification-calibrated. It can also be shown that a classification-calibrated loss function l is also (Fisher) consistent, that is, for any sequence of measurable functions f_n,

$$R_l(f_n) \to R_l(f^*) \Rightarrow R(f_n) \to R(f^*)$$

This essentially tells us that minimizing the hinge loss function is a meaningful way of minimizing the 0/1 loss function. The following theorem shows this.

Theorem 1.2 Let l be a convex and classification-calibrated loss function. Then for all hypothesis in the class,

$$\left|R(\hat{f}) - R(f^*)\right| \leqslant \left|R_l(\hat{f}) - R_l(f^*)\right|$$

Appendix 2. Rademacher complexity

Theorem 2. Let k be a bounded kernel with $sup_x\sqrt{k(x,x)} = \kappa < \infty$ and F be its RKHS. Denote H be a ball with radius r in the RKHS, $H = \{f \in F : \|f\|_F \leqslant r\}$. Then

$$\hat{R}_n(H) \leqslant \frac{r\kappa}{\sqrt{n}}$$

where n is the sample size.

Proof:
$\hat{R}_n(H)$

$= \frac{1}{n}E\left[sup_{f\in H}\sum_{i=1}^n \sigma_i f(x_i)\right]$

$= \frac{1}{n}E\left[sup_{f\in H}\sum_{i=1}^n \sigma_i \langle f, k(\cdot, x_i)\rangle_F\right]$, by the reproducing property of kernels

$= \frac{1}{n}E\left[sup_{f\in H}\langle f, \sum_{i=1}^n \sigma_i k(\cdot, x_i)\rangle_F\right]$, by linearity of inner product

$= \frac{1}{n}E\left[\langle r\frac{\sum_{i=1}^n \sigma_i k(\cdot,x_i)}{\left\|\sum_{i=1}^n \sigma_i k(\cdot,x_i)\right\|}, \sum_{i=1}^n \sigma_i k(\cdot, x_i)\rangle_F\right]$, by Cauchy-Schwartz inequality

$= \frac{r}{n}E\left[\sqrt{\left\|\sum_{i=1}^n \sigma_i k(\cdot, x_i)\right\|^2}\right]$

$\leqslant \frac{r}{n}\sqrt{E\Big[\|\sum_{i=1}^{n}\sigma_i k(\cdot,x_i)\|^2\Big]}$, by Jensen's inequality

$= \frac{r}{n}\sqrt{\sum_{i=1}^{n}\|k(\cdot,x_i)\|^2}$, by properties of the Rademacher random variable

$= \frac{r}{n}\sqrt{\sum_{i=1}^{n}k(x_i,x_i)}$, by the reproducing property of kernels

$\leqslant \frac{r}{n}\sqrt{n}\kappa$

$= \frac{r\kappa}{\sqrt{n}}$

Using the empirical Rademacher complexity $\hat{R}_n(H)$, we can derive the upper bound for the Rademacher complexity by:

$$R_n(H) = E\Big[\hat{R}_n(H)\Big] \leqslant \frac{r\kappa}{\sqrt{n}}$$

Now let's derive the upper bound of Rademacher complexity for each kernel class. Note that the fact that all features are binary (0 or 1) implies that

$$\langle x,x\rangle = \|x\|^2 \leqslant d$$

where d is the feature dimension.

1. Gaussian kernel.

$$\kappa = sup_x\sqrt{k(x,x)} = sup_x\sqrt{exp(-\gamma\|x-x\|^2)} = 1 \Rightarrow R_n(H) \leqslant \frac{r}{\sqrt{n}}$$

2. Sigmoid kernel.

$$\kappa = sup_x\sqrt{k(x,x)} = sup_x\sqrt{tanh(\gamma\langle x,x\rangle)} = \sqrt{tanh(\gamma sup_x\langle x,x\rangle)} = \sqrt{tanh(\gamma d)} \Rightarrow R_n(H) \leqslant \frac{r\sqrt{tanh(\gamma d)}}{\sqrt{n}}$$

3. Homogeneous polynomial kernel.

$$\kappa = sup_x\sqrt{k(x,x)} = sup_x(\gamma\langle x,x\rangle)^{\frac{p}{2}} = (\gamma sup_x\langle x,x\rangle)^{\frac{p}{2}} = (\gamma d)^{\frac{p}{2}} \Rightarrow R_n(H) \leqslant \frac{r(\gamma d)^{\frac{p}{2}}}{\sqrt{n}}$$

Remarks.

- Role of feature dimension d. Note that the Rademacher complexity of Gaussian kernels does not depend on d, but increases exponentially in terms of d for high-degree polynomial kernels. This implies that using high-degree polynomial kernels tend to give large generalization errors when feature dimension is large.
- Sigmoid kernels always have a smaller Rademacher complexity than Gaussian kernels as $\sqrt{tanh(\gamma d)} \leqslant 1$
- When the value γd is small ($<< 1$), polynomial kernels tend to have smaller Rademacher complexity than sigmoid kernels. When the value γd is large (> 1), polynomial kernels tend to have much larger Rademacher complexity than sigmoid kernels.

Appendix 3. Generalization bound based on the Rademacher complexity

Theorem 3. Let k be a bounded kernel with $sup_x\sqrt{k(x,x)} = \kappa < \infty$ and let l be a loss function that is $L - Lipschitz$ with $B = sup_{y\in Y} l(y,0)$. Then with probability $1-\delta$,

$$\left|R(\hat{f}) - R(f^*)\right| \leqslant \left|R_l(\hat{f}) - R_l(f^*)\right| \leqslant sup_{f\in H}\left|R_l(\hat{f}) - R_l(f)\right| \leqslant \frac{2Lr\kappa}{\sqrt{n}} + (B + Lr\kappa)\sqrt{\frac{log(\frac{2}{\delta})}{2n}}$$

Proof:
Let $G = l \circ H = \{(x,y) \to l(y, f(x)) : f \in H\}$, where l is a L-Lipschitz loss function. Then by the Talagrands contraction principle, we have

$$R_n(G) \leqslant LR_n(H) \leqslant \frac{Lr\kappa}{\sqrt{n}}$$

Now we need to find the upper bound of the loss function. Note the following:
$\|f\|_\infty$

$= sup_{x\in X}|f(x)|$

$= sup_{x\in X}|\langle f, k(\cdot,x)\rangle_F|$, using the reproducing property of kernels

$\leqslant sup_{x\in X}\|f\|_F\|k(\cdot,x)\|_F$, by Cauchy-Schwarz of kernels

$\leqslant r\kappa$

Then using the fact that l is L-Lipschitz, we have

$$|l(y, f(x))| \leqslant B + Lr\kappa$$

where $B = sup_{y\in Y} l(y,0)$

Finally using the Rademacher complexity bound (without kernel) of

$$\left|R_l(\hat{f}) - R_l(f^*)\right| \leqslant 2R_n(G) + M\sqrt{\frac{log(\frac{2}{\delta})}{2n}}$$

where $|l| \leqslant M$, we have

$$\left|R_l(\hat{f}) - R_l(f^*)\right| \leqslant \frac{2Lr\kappa}{\sqrt{n}} + (B + Lr\kappa)\sqrt{\frac{log(\frac{2}{\delta})}{2n}}$$

If we are using the hinge loss (which is 1-Lipschitz), then the bound becomes

$$\left|R_l(\hat{f}) - R_l(f^*)\right| \leqslant \frac{2r\kappa}{\sqrt{n}} + (B + r\kappa)\sqrt{\frac{log(\frac{2}{\delta})}{2n}}$$