Evidence Based Model Selection for SVM

Smoliakov Dmitrii

Skoltech

September 20, 2018

Evidence Based Parameters Selection

Suppose that we have probabilistic model Bayesian Formula:

$$\mathbb{P}(\theta|X) = \frac{\mathbb{P}(X|\theta) \cdot \mathbb{P}(\theta|\alpha)}{\mathbb{P}(X|\alpha)} = \frac{\textit{likelihood} \times \textit{prior}}{\textit{evidence}}$$

- $\mathbb{P}(X|\theta)$ – likelihood
- $\mathbb{P}(\theta|\alpha)$ – prior distribution
- α – fixed parameter
- $\mathbb{P}(X|\alpha) = \int_\theta \mathbb{P}(X|\theta) \cdot \mathbb{P}(\theta|\alpha)$ – evidence

Difficulties With Probabilistic SVM

Let's check two class problem for a moment

$$\frac{\|w\|^2}{2} + C\sum_{i=1}^{l} h(y_i[w \cdot \phi(x_i) + b]) \rightarrow \min_{w}$$

Where:

$$h(t) = \max(0, 1 - t)$$

Difficulties With Probabilistic SVM

Let's check two class problem for a moment

$$\frac{\|w\|^2}{2} + C\sum_{i=1}^{l} h(y_i[w \cdot \phi(x_i) + b]) \to \min_w$$

Prior distribution:

$$Q(w) \approx \exp(-\frac{\|w\|^2}{2}) \approx N(0, E)$$

In case of kernel techniques:

$$\theta(x) = w \cdot \phi(x) + b$$

The SVM prior – Gaussian Process

Difficulties With Probabilistic SVM

$$\frac{\|w\|^2}{2} + C \sum_{i=1}^{l} h(y_i[w \cdot \phi(x_i) + b]) \to \min_w$$

Prior distribution:

$$Q(y_i|x_i, w) = k(C) \exp(-Ch(y_i \cdot \theta(x_i)))$$

$k(C)$ is just normalization

$$k(C) = 1/(1 + \exp(-2C))$$

Full likelihood

$$Q(X, y|\theta) = \prod_{i}^{l} Q(y_i|x_i, w)\mathbb{P}(x_i)$$

Difficulties With Probabilistic SVM

$$Q(X, y|\theta) = \prod_{i}^{l} Q(y_i|x_i, w)\mathbb{P}(x_i)$$

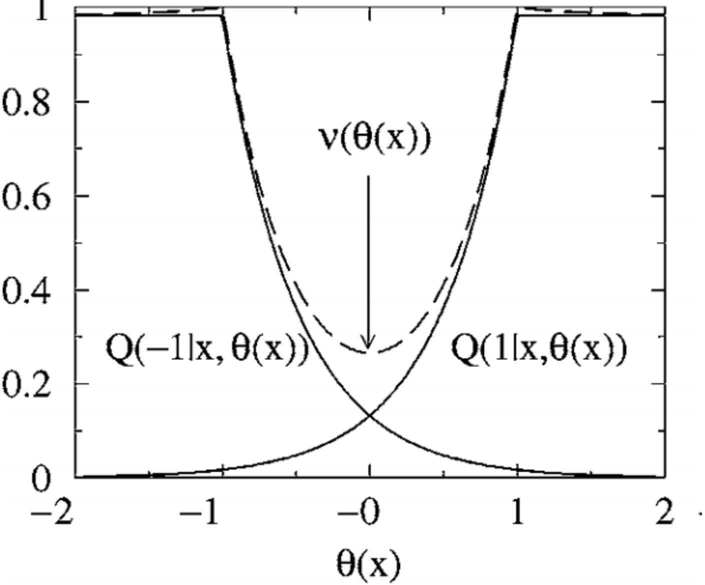
Take a look at a single point:

$$\nu(\theta(x)) = Q(1|x,\theta) + Q(-1|x,\theta) = \\ k(C)\left(\exp(-Cl(\theta(x))) + \exp(-Cl(-\theta(x)))\right) \leq 1$$

Sum of all possible sets is less than one

$$\int_{X,y} Q(X, y|\theta) = \left(\int_{x} Q(x)\nu(\theta(x))\right)^{l} \leq 1$$

Difficulties With Probabilistic SVM



Difficulties With Probabilistic SVM

In the paper "Bayesian Methods for Support Vector Machines: Evidence and Predictive Class Probabilities. Authors proposed to add special normalize coefficient.

$$\mathbb{P}(X, y, \theta) = Q(X, y|\theta)Q(\theta)/N(X, y)$$

Where:

$$N(\theta) = \int_x Q(x)\nu(\theta(x))$$

$$N(D) = \int d\theta Q(\theta)N^n(\theta)$$

Probabilistics SVM

Data is produced by this mechanism

1. Generate θ from GP prior
2. Sample x from $Q(x)$
3. Assign labels with probabilities $Q(y|x, \theta)$
4. With probability $1 - \nu(\theta(x))$ generate "I don't know" class
5. If single "I don't know" was generated – restart procedure

$|\theta(x)|$ is small inside the gap this leads to a bigger margin.

One Class SVM

One of the possible options of building a model of the normal condition is One Class SVM.

We have:

- Points $X_1, \ldots, X_l \subset \mathbb{R}^m$
- Mapping $\phi : \mathbb{R}^m \to \mathbb{H}_\phi$

We want:

- Separate points from coordinate origin in $\mathbb{H}_\phi$

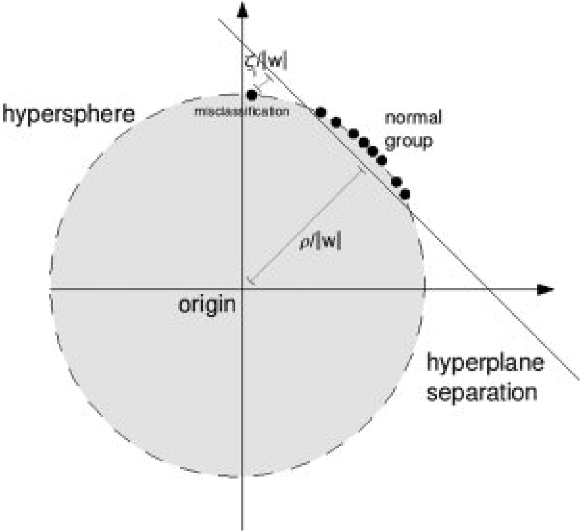
Optimization Problem

$$\frac{\nu l}{2}\|w\|^2 - \rho\nu l + \sum_{i=1}^{l}\xi_i \to \min_{w,\rho,\xi}$$

$$(w \cdot \phi(X_i)) \geq \rho - \xi_i$$

$$\xi_i \geq 0$$

Intuition Illustration



Problems

No free lunch

1. No explicit labeling
2. No proper validation techniques
3. Difficult to select parameters

One Class SVM

$$\frac{\|w\|^2}{2} + C\sum_{i=1}^{l} h([w \cdot \phi(x_i) + b]) \to \min_{w}$$

1. Generate θ from GP prior
2. Sample x from $Q(x)$
3. With probability $1 - \prod_{i=1}^{l}(1 - \nu(\theta(x_i)))$ repeat

We will get data from some distribution $Q(x)$ but based on decision function $\theta(x)$ some of elements are less probable and we reject the whole set.

Evidence Calculation

We are going to look at this values:
Probability of X:

$$\mathbb{P}(X) = \frac{N(x)}{N} Q(X)$$

Probability not to reject given X

$$\mathbb{P}(1|X) = Q(1|X)N(X)$$

Calculating $Q(1|X)$

$Q(1|X)$ – multidimensional integral. We will use Laplace approximation

$$Q(1|X) = k^n(C)\int \frac{d\theta}{\sqrt{2\pi K}} \exp\left(-\frac{1}{2}\sum_{i,j}\theta_i(K^{-1})_{i,j}\theta_j - C\sum_i h(\theta_i)\right) =$$

$$\underbrace{k^n(C)\exp\left(-\frac{1}{2}\sum_{i,j}\theta_i^*(K^{-1})\theta_j^* - C\sum_i h(\theta_i^*)\right)}_{Q^*(1|X)}$$

$$\times \int \frac{d\Delta\theta}{\sqrt{2\pi|K|}} \exp -\frac{1}{2}\sum_{i,j}\Delta\theta_i(K^{-1})\Delta\theta_j -$$

$$\sum_{i,j}\Delta\theta(K^{-1})_{i,j}\theta_j^* - C\sum_i \Delta\theta_i H(\Delta\theta_j)$$

Calculating $Q(1|X)$

Finally:

$$Q(1|X) \approx Q^*(1|X) \frac{1}{\sqrt{2\pi|K|}} \prod_i \left(\frac{1}{C - \alpha_i} + \frac{1}{\alpha_i} \right)$$

Where:

$$\alpha_i = \sum_j K^{-1}_{i,j} \theta^*_j$$

Calculating $N(X)$

Authors of initial paper calculated coefficient $N(X)$ numerically.

Summary

In the end it's better than nothing

- It's hard to select good parameters in anomaly detection
- Evidence maximization gives an efficient framework for model selection
- Vanilla One Class SVM doesn't have proper probability model
- Introducing additional "I don't know" class allows to build probability model
- Unfortunately it contains numerical calculation of multidimensional integral