Auto LS-SVM

Laurent Sorber

25 August 2017

1 Auto LS-SVM

1.1 Goals

Auto LS-SVM is a Least Squares Support Vector Machine (LS-SVM) that is improved in a number of ways *to enable effortless construction of LS-SVMs that generalise well to new data*:

1. A next-gen regularization term that penalizes model complexity directly:
   (a) Regression: the objective is a tradeoff between fit on training data and a direct estimation of model complexity (instead of a proxy like the L2 norm).
   (b) Classification: the objective is a mix between fit on the training data, model complexity (next-gen term) and maximizing the margin (L2 norm). While maximizing the margin is important, minimizing the decision boundary complexity may significantly improve generalization performance.
2. The optimisation problem is "fully normalised" to achieve these subgoals:
   (a) Hyperparameters are easily interpreted.
   (b) Default hyperparameter values (say 1.0) will give good results.
   (c) Adding or removing rows (observations) or columns (features) has a minimal influence of the choice of the hyperparameter values. Current hyperparameter definitions are not invariant to changes in these two dimensions.

3. The least squares formulation allows for a cheap closed-form expression for the leave-one-out error. This enables a powerful way to search for the optimal hyperparameters.

4. In a later stage, perhaps a large scale implementation too (e.g., based on Random Kitchen Sinks).

1.2 A next-gen regularization term

We begin with an observation x, and dual variables α:

$$x, x_i \in R^{d\times 1} \tag{1}$$

$$\alpha \in R^{n\times 1} \tag{2}$$

$$X := [x_1 \ \dots \ x_n] \in R^{d\times n} \tag{3}$$

We choose an RBF kernel in the same format as `sklearn`:

$$k(x, y) := \exp(-\gamma \|x - y\|^2) \in R \tag{4}$$

$$K(x) := [k(x, x_1) \ \dots \ k(x, x_n)] \in R^{1\times n} \tag{5}$$

The gradient of the kernel is given by:

$$\frac{dk}{dx} = -2\gamma k(x, y) \cdot (x - y) \tag{6}$$

$$\nabla K := \frac{dK}{dx} = -2\gamma \left(x \cdot 1_{1\times n} - X\right) \cdot \mathrm{diag}(K(x)) \tag{7}$$

The SVM model is of the form:

$$f(x) := K(x) \cdot \alpha \tag{8}$$

$$\min_{\alpha} \|y - f(x)\|^2 + \mu \cdot \mathrm{nextgen} + \nu \cdot \|\alpha\|^2 \tag{9}$$

The normal on the prediction surface is:

$$n := \begin{bmatrix} \nabla K \cdot \alpha \\ -1 \end{bmatrix} \tag{10}$$

And so the norm of the normal vector is:

$$\|n\|^2 = \alpha^T \cdot (\nabla K)^T (\nabla K) \cdot \alpha + 1 \tag{11}$$

$$= 1 + 4\gamma^2 \cdot \alpha^T \cdot \mathrm{diag}(K(x)) \cdot \left(\|x\|^2 - 1_{n\times 1} \cdot x^T \cdot X - X^T \cdot x \cdot 1_{1\times n} + X^T X\right) \cdot \mathrm{diag}(K(x)) \cdot \alpha \tag{12}$$

$$= 1 + 4\gamma^2 \cdot \alpha^T \cdot \left[k(x, x_i) k(x, x_j) \left(\|x\|^2 - x_i^T \cdot x - x_j^T \cdot x + x_i^T \cdot x_j\right)\right] \cdot \alpha \tag{13}$$

We can integrate $\|n\|$ over the prediction surface to obtain its d-volume, but then we need a definite integral with finite bounds. Instead, we can integrate $\|\nabla K \cdot \alpha\|$ (the norm of the gradient of the prediction surface) and use infinite bounds, which has the effect of simplifying the integral.

1.2.1 Derivation of the regularization term

Let's integrate each of the three types of terms.

Term of the form $k(x, x_i)k(x, x_j)x_i^T \cdot x_j$

First, let's integrate out the p-th dimension of x:

$$
\begin{aligned}
I_{ij}^{(3,p)} &= \int_{-\infty}^{\infty} k(x, x_i)k(x, x_j)x_i^T \cdot x_j dx^{(p)} && (14)\\
&= k(x^{(/p)}, x_i^{(/p)})k(x^{(/p)}, x_j^{(/p)})x_i^T \cdot x_j \int_{-\infty}^{\infty} \exp(-\gamma(x^{(p)} - x_i^{(p)})^2 - \gamma(x^{(p)} - x_j^{(p)})^2)dx^{(p)} && (15)\\
&= C_{ij} \int_{-\infty}^{\infty} \exp(-\gamma(x^{(p)} - x_i^{(p)})^2 - \gamma(x^{(p)} - x_j^{(p)})^2)dx^{(p)} && (16)\\
&= C_{ij} \int_{-\infty}^{\infty} \exp\left(-2\gamma\left(x^{(p)} - \frac{x_i^{(p)} + x_j^{(p)}}{2}\right)^2 - \frac{\gamma}{2}\left(x_i^{(p)} - x_j^{(p)}\right)^2\right) dx^{(p)} && (17)\\
&= C_{ij}\sqrt{\frac{\pi}{2\gamma}} \exp\left(-\frac{\gamma}{2}\left(x_i^{(p)} - x_j^{(p)}\right)^2\right) && (18)
\end{aligned}
$$

This means that after integrating out all d dimensions, we get:

$$
I_{ij}^{(3)} = x_i^T \cdot x_j \left(\frac{\pi}{2\gamma}\right)^{\frac{d}{2}} \sqrt{k(x_i, x_j)} \quad (19)
$$

Term of the form $k(x, x_i)k(x, x_j)x_i^T \cdot x$

First, let's integrate out the p-th dimension ($p \neq q$) of x:

$$I_{ij}^{(2,p,i)} = \int_{-\infty}^{\infty} k(x,x_i)k(x,x_j)x_i^T \cdot x dx^{(p)} \tag{20}$$

$$= \sum_{q=1}^{d} \int_{-\infty}^{\infty} k(x,x_i)k(x,x_j)x_i^{(q)}x^{(q)}dx^{(p)} \tag{21}$$

$$= \sum_{q=1}^{d} k(x^{(/p)},x_i^{(/p)})k(x^{(/p)},x_j^{(/p)})x_i^{(q)}x^{(q)}\sqrt{\frac{\pi}{2\gamma}}\exp\left(-\frac{\gamma}{2}\left(x_i^{(p)}-x_j^{(p)}\right)^2\right) \tag{22}$$

Next, we integrate out all other dimensions:

$$I_{ij}^{(2,i)} = \sum_{q=1}^{d}\left(\frac{\pi}{2\gamma}\right)^{\frac{d-1}{2}}\sqrt{k(x_i^{(/q)},x_j^{(/q)})}\int_{-\infty}^{\infty} k\left(x^{(q)},x_i^{(q)}\right)k\left(x^{(q)},x_j^{(q)}\right)x_i^{(q)}x^{(q)}dx^{(q)} \tag{23}$$

$$= \left(\frac{\pi}{2\gamma}\right)^{\frac{d}{2}}\sqrt{k(x_i,x_j)}\sum_{q=1}^{d} x_i^{(q)}\frac{x_i^{(q)}+x_j^{(q)}}{2} \tag{24}$$

$$= \left(\frac{\pi}{2\gamma}\right)^{\frac{d}{2}}\sqrt{k(x_i,x_j)}x_i\cdot(x_i+x_j)/2 \tag{25}$$

Term of the form $k(x,x_i)k(x,x_j)\|x\|^2$
First, let's integrate out the p-th dimension ($p \neq q$) of x:

$$I_{ij}^{(1,p)} = \int_{-\infty}^{\infty} k(x,x_i)k(x,x_j)\|x\|^2 dx^{(p)} \tag{26}$$

$$= \sum_{q=1}^{d}\int_{-\infty}^{\infty} k(x,x_i)k(x,x_j)x^{(q)2}dx^{(p)} \tag{27}$$

$$= \sum_{q=1}^{d} k(x^{(/p)},x_i^{(/p)})k(x^{(/p)},x_j^{(/p)})x^{(q)2}\sqrt{\frac{\pi}{2\gamma}}\exp\left(-\frac{\gamma}{2}\left(x_i^{(p)}-x_j^{(p)}\right)^2\right) \tag{28}$$

Next, we integrate out all other dimensions:

$$I_{ij}^{(1)} = \sum_{q=1}^{d} \left(\frac{\pi}{2\gamma}\right)^{\frac{d-1}{2}} \sqrt{k(x_i^{(/q)}, x_j^{(/q)})} \int_{-\infty}^{\infty} k\left(x^{(q)}, x_i^{(q)}\right) k\left(x^{(q)}, x_j^{(q)}\right) x^{(q)2} dx^{(q)} \tag{29}$$

$$= \left(\frac{\pi}{2\gamma}\right)^{\frac{d}{2}} \sqrt{k(x_i, x_j)} \sum_{q=1}^{d} \frac{1}{4}\left((x_i^{(q)} + x_j^{(q)})^2 + \frac{1}{\gamma}\right) \tag{30}$$

$$= \left(\frac{\pi}{2\gamma}\right)^{\frac{d}{2}} \sqrt{k(x_i, x_j)} \frac{1}{4}\left(\|x_i + x_j\|^2 + \frac{d}{\gamma}\right) \tag{31}$$

Summing the terms up:

$$I = I_{ij}^{(1)} - I_{ij}^{(2,i)} - I_{ij}^{(2,j)} + I_{ij}^{(3)} \tag{32}$$

$$= \left(\frac{\pi}{2\gamma}\right)^{\frac{d}{2}} \sqrt{k(x_i, x_j)} \frac{1}{4}\left(\|x_i + x_j\|^2 + \frac{d}{\gamma}\right) \tag{33}$$

$$- \left(\frac{\pi}{2\gamma}\right)^{\frac{d}{2}} \sqrt{k(x_i, x_j)} x_i \cdot (x_i + x_j)/2 \tag{34}$$

$$- \left(\frac{\pi}{2\gamma}\right)^{\frac{d}{2}} \sqrt{k(x_i, x_j)} x_j \cdot (x_i + x_j)/2 \tag{35}$$

$$+ \left(\frac{\pi}{2\gamma}\right)^{\frac{d}{2}} \sqrt{k(x_i, x_j)} x_i^T \cdot x_j \tag{36}$$

$$= \left(\frac{\pi}{2\gamma}\right)^{\frac{d}{2}} \sqrt{k(x_i, x_j)} \left(\frac{1}{4}\left(\|x_i + x_j\|^2 + \frac{d}{\gamma}\right) - x_i \cdot (x_i + x_j)/2 - x_j \cdot (x_i + x_j)/2 + x_i^T \cdot x_j\right) \tag{37}$$

$$= \left(\frac{\pi}{2\gamma}\right)^{\frac{d}{2}} \sqrt{k(x_i, x_j)} \left(\frac{1}{4}\left(\|x_i + x_j\|^2 + \frac{d}{\gamma}\right) - \frac{1}{2}\|x_i\|^2 - \frac{1}{2}\|x_j\|^2\right) \tag{38}$$

$$= \frac{1}{4}\left(\frac{\pi}{2\gamma}\right)^{\frac{d}{2}} \sqrt{k(x_i, x_j)} \left(\frac{d}{\gamma} - \|x_i - x_j\|^2\right) \tag{39}$$

1.2.2 The resulting formula

$$\int \|\nabla f\|^2 = \int \alpha^T \cdot (\nabla K)^T (\nabla K) \cdot \alpha \tag{40}$$

$$= \gamma^2 \left(\frac{\pi}{2\gamma}\right)^{\frac{d}{2}} \cdot \alpha^T \cdot \left[\sqrt{k(x_i, x_j)} \left(\tfrac{d}{\gamma} - \|x_i - x_j\|^2\right)\right] \cdot \alpha \tag{41}$$

$$\propto \gamma^{-\frac{d}{2}} \cdot \alpha^T \cdot \left[\sqrt{k(x_i, x_j)} \left(d\gamma - \gamma^2 \|x_i - x_j\|^2\right)\right] \cdot \alpha \tag{42}$$

1.3 Normalization of the full problem

We want to achieve a number of things:

1. The solution should be close to invariant under addition or removal of columns.
2. The solution should be close to invariant under addition or removal of rows.
3. A default value of $\gamma = 1.0$ should be a good starting value for data sets of any size.
4. A default value of $\mu = 1.0$ should be a good starting value for data sets of any size.

Reduction of sensitivity of the regularization term on d

Note: the below are just a few initial thoughts to achieve the above goals. Needs further exploration, but pretty certain it can be done.

First, we fix $d = 2$. This way, it is as if we integrated only two dimensions, making the exponential effect of γ less problematic.

$$\alpha^T \cdot \left[\sqrt{k(x_i, x_j)} \left(2 - \gamma \|x_i - x_j\|^2\right)\right] \cdot \alpha \tag{43}$$

Second, we replace γ with $\frac{\gamma}{d}$ in the kernel function k. This makes it so that if you only have one feature ($d = 1$) and you add a copy of this feature to the feature matrix ($d = 2$), you can keep the same value for γ and end up with an identical model.

$$\alpha^T \cdot \left[\sqrt{k(x_i, x_j)}\left(2 - \tfrac{\gamma}{d}\|x_i - x_j\|^2\right)\right] \cdot \alpha \tag{44}$$

A counterargument to this is that adding an all-zero feature would require a different value of γ to reach optimality.