Exercises
Machine Learning
HS 2016

Machine Learning Laboratory
Dept. of Computer Science, ETH Zürich
Prof. Joachim M. Buhmann
Web https://ml2.inf.ethz.ch/courses/ml/

Series 4, Nov 21th, 2015
(Non-linear SVMs & Kernels)

For questions 1,2,3,4 : **Jiadong Guo**
jguo@student.ethz.ch
For questions 5,6 : **Junyao Zhao**
zhaoju@student.ethz.ch

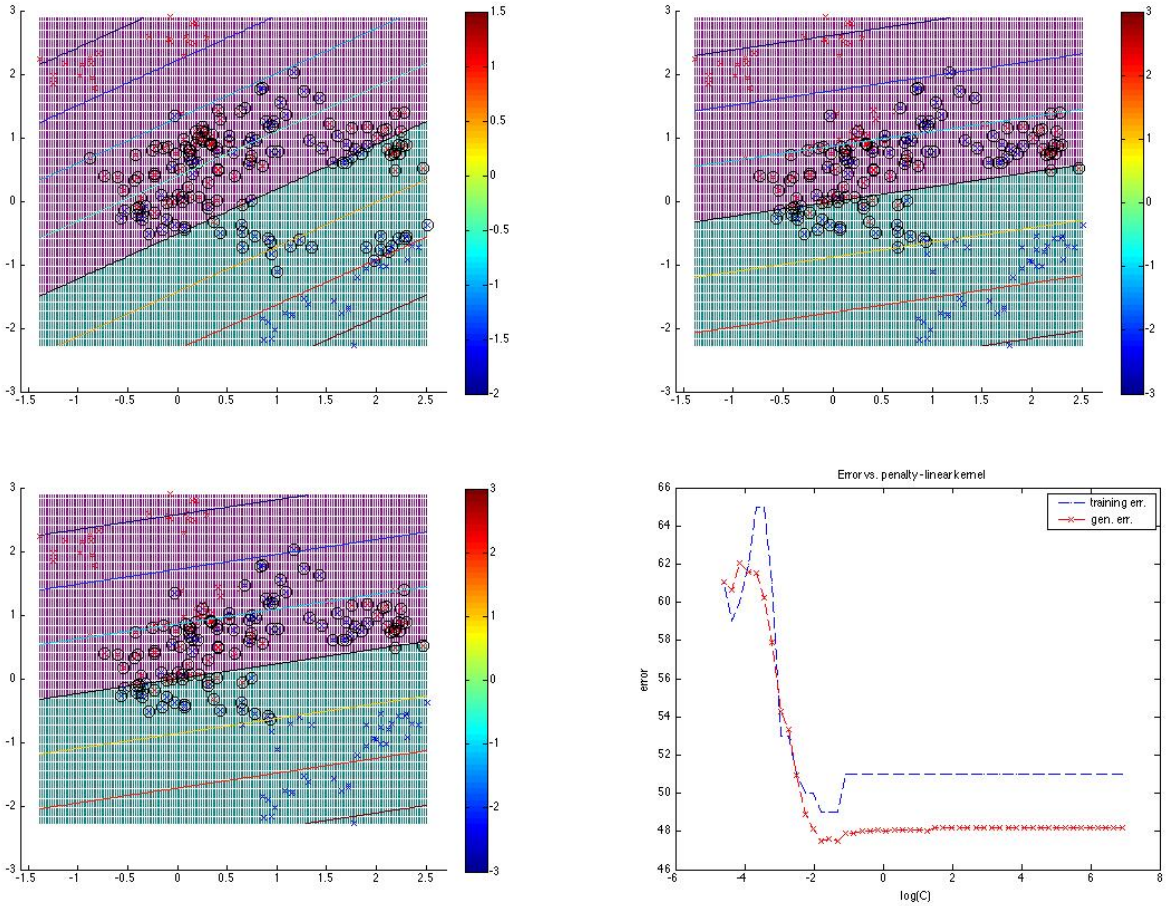
Problem 1 (Overfitting with non-linear SVMs):

The file overfitting_nonlin_SVMs.m, provided in the course website, contains the code necessary to answer the exercise.

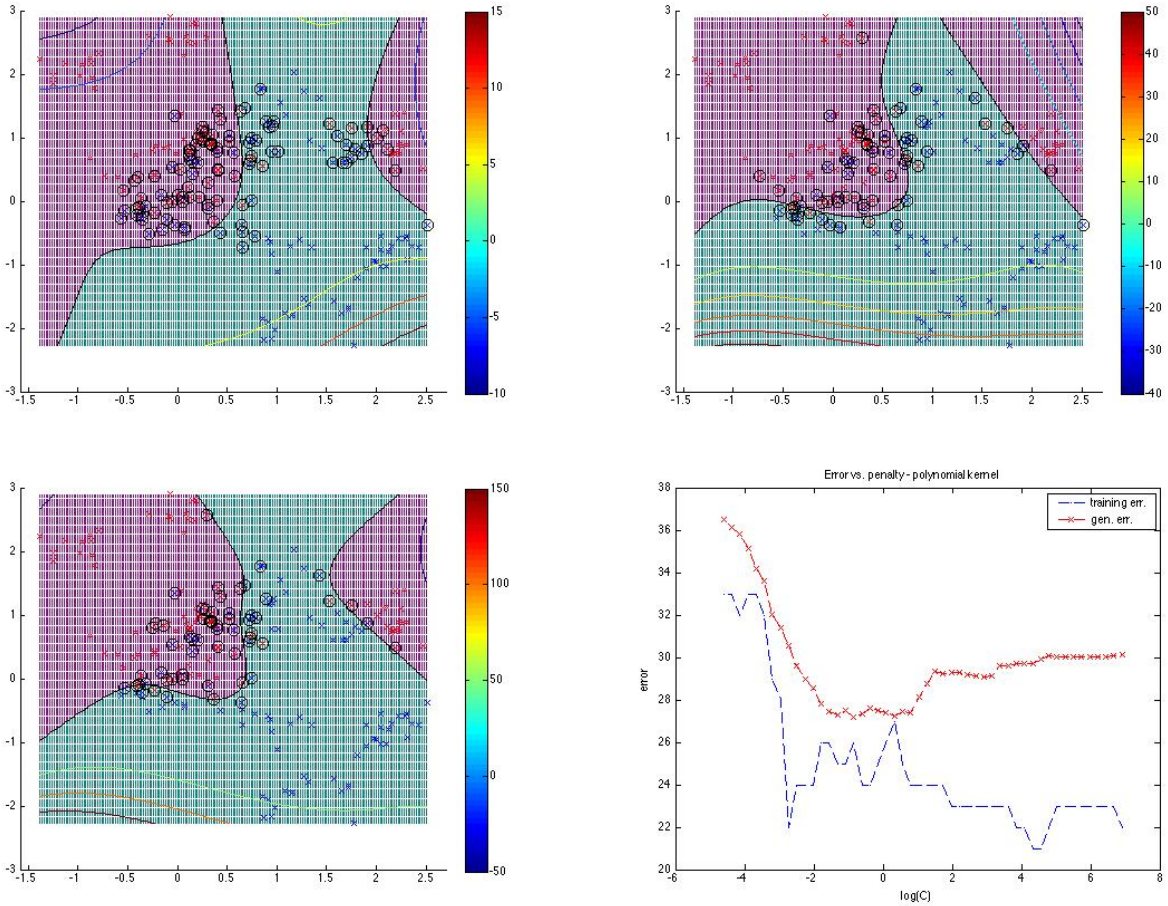
We make the following observations regarding the relationship between the penalty parameter C, the type of kernel $k(\mathbf{x}, \mathbf{y})$ and the bias-variance trade-off as it applies to SVMs:

- A kernel $k(\mathbf{x}, \mathbf{y})$ that corresponds to a large feature space increases linear separabilty of the data in the enlarged feature space. If all margin variables ξ_i are equal to zero then the data has been separated perfectly by the hyperplane in the enlarged feature space. This probably corresponds to overfitting unless there is a good reason to believe that classes are linearly separable **for all possible realizations of training data**.
- A larger value of C in enlarged feature space will discourage data points inside the margin ($\xi_i > 0$) and lead to an overfitt, *wiggly* boundary which *tries* to correctly classify **all** data points. This can be seen in the primal soft-SVM formulation: if C is large, the $C \sum_{i=1}^{n} \xi_i$ term dominates the $\frac{1}{2}\mathbf{w}^T\mathbf{w}$ term and the optimization algorithm is *forced* to drive ξ_i's to zero.

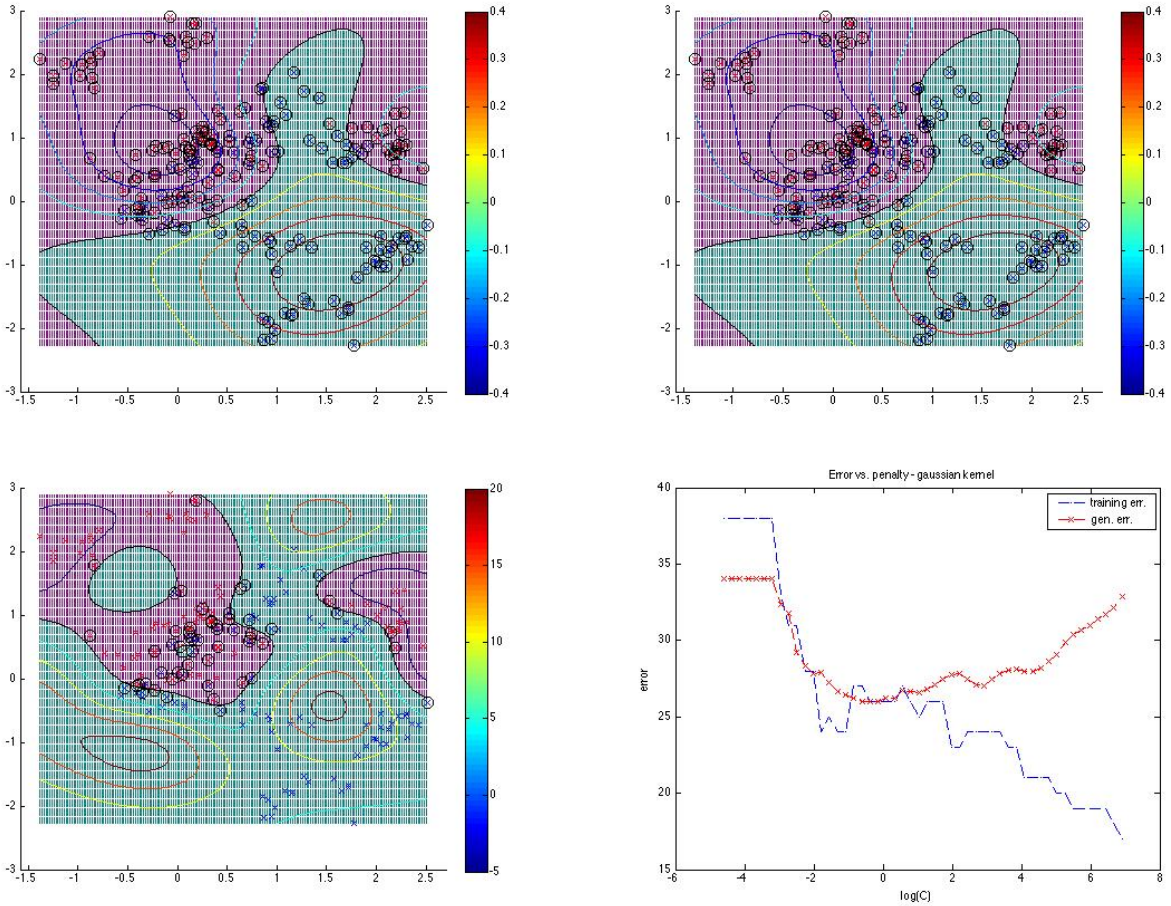
These observations can be corroborated using the graphs generated in exercise 1.



The figure shows the SVM *linear* models for $C = 0.02, 1, 1000$ and the training and generalization error for a range of Cs. For low values of C there is improvement in the error as C increases, resulting from a rotation of the hyperplane. Although, increasing the value of C does not lead to overfitting, the overall level of generalization error is quite poor (40-60%). The model has low flexibility, which makes it robust to overfitting but also very biased.



The figure shows the SVM *polynomial* models for $C = 0.02, 1, 1000$ and the training and generalization error for a range of Cs. This model has much greater flexibility than the linear one, which permits it to drive the bias down as can be seen by lower levels of generalization error(28-36%). For large, values of C we begin to see more *wiggly* decision boundaries which result in overfitting.



The figure shows the SVM *RBF* models for $C = 0.02, 1, 1000$ and the training and generalization error for a range of Cs. This model also has high flexibility which permits it to drive the bias down. In fact, not surprisingly since the simulated data is a mixture of gaussians, this model achieved the lowest generalization error(26-34%). For large values of C we also begin to see more *wiggly* decision boundaries which result in overfitting.

Problem 2 (Identifying Kernel Functions):

1. $k(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{d} h(\frac{x_i - c}{a}) h(\frac{y_i - c}{a})$ where $h(x) = \cos(1.75x)\exp(-x^2/2)$.

   Since $k(\mathbf{x}, \mathbf{y}) = f(\mathbf{x})f(\mathbf{y})$ where $f(\mathbf{x}) = \prod_{i=1}^{d} h(\frac{x_i - c}{a})$ it follows that the Gram matrix of k can be written as $\bar{f}\bar{f}^T$ where $\bar{f} = [f(\mathbf{x}_1)f(\mathbf{x}_2)\dots f(\mathbf{x}_n)]$ for any $\{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset \mathbb{R}^d$. Therefore k has a p.s.d Gram matrix (formed using any finite subset of $\mathbb{R}^d$) and hence is a kernel function.

2. $k(\mathbf{x}, \mathbf{y}) = -\log_2(\|\mathbf{x} - \mathbf{y}\| + 1)$.

   Choose $\mathbf{x}$ and $\mathbf{y}$ such that $\|\mathbf{x} - \mathbf{y}\| = 3$. Then the corresponding Gram matrix of k formed using these two points is:

   $$\mathbf{K} = \begin{pmatrix} 0 & -2 \\ -2 & 0 \end{pmatrix}.$$

   The eigenvalues of $\mathbf{K}$ are ± 2 implying $\mathbf{K}$ is not p.s.d and hence k is not a kernel function.

3. $k(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{d}(x_i - y_i)$.
   Clearly $k(\mathbf{x},\mathbf{y}) = -k(\mathbf{y},\mathbf{x})$ and so is not symmetric. Hence k is not a kernel function.

4.* $k(X,Y) = 2^{|X\cap Y|}$ for $X, Y \subseteq \Omega$ where Ω is a finite set and $|\cdot|$ denotes cardinality.
   Assume without loss of generality that $\Omega = \{1,2,3,\ldots,m\}$ so $|\Omega| = m$. Denote each $X \subseteq \Omega$ by a m-bit string $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ where $x_i = 1$ if $i \in X$ and 0 otherwise. Similarly represent any other $Y \subseteq \Omega$ as $\mathbf{y} = (y_1, y_2, \ldots, y_m)$. It follows that

$$k(X,Y) = 2^{|X\cap Y|} = 2^{\sum_{i=1}^{m} x_i y_i} = 2^{\langle \mathbf{x},\mathbf{y}\rangle} = \exp(\langle \mathbf{x},\mathbf{y}\rangle \ln 2).$$

   Note that $\ln 2 > 0$ and thus $k'(\mathbf{x},\mathbf{y}) = \langle \mathbf{x},\mathbf{y}\rangle \ln 2$ is a kernel. Lastly we know that $\exp(k'(\mathbf{x},\mathbf{y}))$ is a kernel whenever k' is a kernel. Hence k is a kernel.

Problem 3 (Normalized and Gaussian kernels):

1. We have $k'(\mathbf{x},\mathbf{y}) = \frac{k(\mathbf{x},\mathbf{y})}{\sqrt{k(\mathbf{x},\mathbf{x})k(\mathbf{y},\mathbf{y})}}$. Let us first see that k' is a kernel. Indeed we have:

$$k'(\mathbf{x},\mathbf{y}) = \frac{k(\mathbf{x},\mathbf{y})}{\sqrt{k(\mathbf{x},\mathbf{x})k(\mathbf{y},\mathbf{y})}} = \langle \frac{\phi(\mathbf{x})}{\sqrt{k(\mathbf{x},\mathbf{x})}}, \frac{\phi(\mathbf{y})}{\sqrt{k(\mathbf{y},\mathbf{y})}} \rangle = \langle \phi'(\mathbf{x}), \phi'(\mathbf{y}) \rangle.$$

   In the second equality we used the fact that $k(\mathbf{x},\mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y})\rangle$ with ϕ being the the feature map associated with k. In the third equality we simply obtain $\phi'(\cdot) = \frac{\phi(\cdot)}{\sqrt{k(\cdot,\cdot)}}$. Hence k' is a kernel. To show that k' takes values in $[-1,1]$ we will show that: $k^2(\mathbf{x},\mathbf{y}) \leq k(\mathbf{x},\mathbf{x})k(\mathbf{y},\mathbf{y})$. Indeed consider the 2×2 gram matrix $\mathbf{K}$ formed using any $\mathbf{x},\mathbf{y} \in \mathcal{X}$:

$$\mathbf{K} = \begin{pmatrix} k(\mathbf{x},\mathbf{x}) & k(\mathbf{x},\mathbf{y}) \\ k(\mathbf{y},\mathbf{x}) & k(\mathbf{y},\mathbf{y}) \end{pmatrix}.$$

   Since $\mathbf{K}$ is p.s.d and using the fact : $k(\mathbf{x},\mathbf{y}) = k(\mathbf{y},\mathbf{x})$ (symmetry of k) we have that

$$\det(\mathbf{K}) \geq 0 \Leftrightarrow k(\mathbf{x},\mathbf{x})k(\mathbf{y},\mathbf{y}) - k^2(\mathbf{x},\mathbf{y}) \geq 0.$$

2. We need to show that $\tilde{k}(\mathbf{x},\mathbf{y}) = f(\mathbf{x})k(\mathbf{x},\mathbf{y})f(\mathbf{y})$ is a kernel function. Using the fact that $k(\mathbf{x},\mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y})\rangle$ we have $\tilde{k}(\mathbf{x},\mathbf{y}) = \langle f(\mathbf{x})\phi(\mathbf{x}), f(\mathbf{y})\phi(\mathbf{y})\rangle = \langle \phi'(\mathbf{x}), \phi'(\mathbf{y})\rangle$ where $\phi'(\cdot) = f(\cdot)\phi(\cdot)$. Hence $\tilde{k}$ is a kernel function.

3. We have that:

$$\exp(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{h^2}) = \underbrace{\exp\left(-\frac{\|\mathbf{x}\|^2}{h^2}\right)}_{f(\mathbf{x})} \exp\left(2\frac{\langle \mathbf{x},\mathbf{y}\rangle}{h^2}\right) \underbrace{\exp\left(-\frac{\|\mathbf{y}\|^2}{h^2}\right)}_{f(\mathbf{y})}.$$

   Observe that $2\frac{\langle \mathbf{x},\mathbf{y}\rangle}{h^2}$ is a kernel function and so $\exp(2\frac{\langle \mathbf{x},\mathbf{y}\rangle}{h^2})$ is also a kernel function. Hence it follows easily by using (2) that $\exp(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{h^2})$ is a kernel function.

Problem 4 (Support Vector Regression):

1. The epsilon sensitivity function is not differentiable when the error $f(\mathbf{x}) - y$ equals to ϵ, which means when the point lies exactly on one of the two boundaries. Thus, we can introduce slack variables ξ and ξ^* to account for errors in points that lie outside the ϵ tube as follows.

$$y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle - \epsilon \leq \xi_i \tag{1}$$
$$\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i - \epsilon \leq \xi_i^* \tag{2}$$
$$\xi, \xi^* \geq 0. \quad i = 1, \cdots, n \tag{3}$$

Adding two slack variables for each data point, now we have added in total 2n slack variables. Now the primal form can be written down as :

$$\min_{\mathbf{w} \in R^m, \xi \in R^n, \xi^* \in R^n} \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \tag{4}$$

with the additional inequality constraints from inequalities 1 to 3.

2. There are 4 inequality constraints for every data point, thus we add 4 additional Lagrange multipliers($\alpha_i, \alpha_i^*, \beta_i, \beta_i^*$) correspondingly.

$$L = L(\mathbf{w}, \xi, \xi^*, \alpha, \alpha^*, \beta, \beta^* := \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \tag{5}$$
$$- \sum_{i=1}^{n} (\beta_i \xi_i + \beta_i^* \xi_i^*) \tag{6}$$
$$- \sum_{i=1}^{n} \alpha_i (\epsilon + \xi_i - y_i + \langle \mathbf{w}, \mathbf{x}_i \rangle) \tag{7}$$
$$- \sum_{i=1}^{n} \alpha_i^* (\epsilon + \xi_i^* + y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle) \tag{8}$$

with positivity constraints on the Lagrangian multipliers,

$$\alpha_i, \alpha_i^*, \beta_i, \beta_i^* \geq 0, (i = 1, \cdots, n) \tag{9}$$

3. Taking derivative of L w.r.t the primal variables $(\mathbf{w}, \xi_i, \xi_i^*)$, one gets

$$\partial_{\mathbf{w}} L = \mathbf{w} - \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) x_i = 0 \tag{10}$$
$$\partial_{\xi_i} L = C - \alpha_i - \beta = 0 \tag{11}$$
$$\partial_{\xi_i^*} L = C - \alpha_i^* - \beta_i^* = 0 \tag{12}$$

from the last two equations one get

$$0 \leq \beta_i = C - \alpha_i \tag{13}$$
$$0 \leq \beta_i^* = C - \alpha_i^* \tag{14}$$

Substituting the result back into the Lagrangian equation 15:

$$L = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{i=1}^{n}(\beta_i\xi_i + \beta_i^*\xi_i^*) - \sum_{i=1}^{n}\alpha_i(\epsilon + \xi_i - y_i + \langle \mathbf{w}, \mathbf{x}_i\rangle) - \sum_{i=1}^{n}\alpha_i^*(\epsilon + \xi_i^* + y_i - \langle \mathbf{w}, \mathbf{x}_i\rangle) \tag{15}$$

$$= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle \mathbf{x}_i, \mathbf{x}_j\rangle + \sum_{i=1}^{n}\xi_i\underbrace{(C - \beta_i - \alpha_i)}_{0} + \sum_{i=1}^{n}\xi_i^*\underbrace{(C - \beta_i^* - \alpha_i^*)}_{0} \tag{16}$$

$$- \epsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n}y_i(\alpha_i - \alpha_i^*) - \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\underbrace{(\sum_{j=1}^{n}(\alpha_j - \alpha_j^*)\langle \mathbf{x}_i, \mathbf{x}_j\rangle)}_{\langle \mathbf{w}, \mathbf{x}_i\rangle} \tag{17}$$

$$= -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle \mathbf{x}_i, \mathbf{x}_j\rangle - \epsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n}y_i(\alpha_i - \alpha_i^*) \tag{18}$$

So that we reach the corresponding dual problem as follows:

$$\max_{\alpha,\alpha^*} -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle \mathbf{x}_i, \mathbf{x}_j\rangle - \epsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n}y_i(\alpha_i - \alpha_i^*) \tag{19}$$

$$subject\ to \quad \alpha_i, \alpha_i^* \in [0, C] \tag{20}$$

The inequality constraints follow from the equations 14 and 9. The formulation only includes the dual variables α_i and α_i^*. Together with the inequality constraints from the primal formulation, one can see that the KKT conditions are indeed satisfied in our optimization. *

4. Recall the KKT complementary slackness conditions. In the optimal solutions of the dual one has:

$$\alpha_i(\epsilon + \xi_i - y_i + \langle \mathbf{w}, \mathbf{x}_i\rangle) = 0 \tag{21}$$
$$\alpha_i^*(\epsilon + \xi_i^* + y_i - \langle \mathbf{w}, \mathbf{x}_i\rangle) = 0 \tag{22}$$
$$\beta_i\xi_i = 0 \tag{23}$$
$$\beta_i^*\xi_i^* = 0 \tag{24}$$
$$\tag{25}$$

for all i = 1, $\cdots$ n.

Equation 22 indicates that if $\alpha_i > 0$, then $\epsilon + \xi_i - y_i + \langle \mathbf{w}, \mathbf{x}_i\rangle = 0$. So similar to SVM, the points are selected as support vectors, when its corresponding Lagrangian variable α_i is greater than zero.

Now for one support vector, if $\xi = 0$, then it implies that $\mathbf{x}_i$ is on the border of the ϵ tube, therefore $\mathbf{x}_i$ is a margin support vector. If $\xi_i > 0$, then it means the point lies outside of the tube. These $\mathbf{x}_i$ correspond to an non/margin support vector. Similar reasoning holds for ξ_i^* and α_i^*.

5. Prediction in primal:

$$y_{new} = f(X) = \mathbf{w}^T X \tag{26}$$

*To read more about how the dual formulation is derived and what KKT conditions are, the machine learning class from stanford has some resources available under *http://cs229.stanford.edu/notes/cs229-notes3.pdf*

And the corresponding dual by substituting the vector $\mathbf{w}$:

$$f(X) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\langle \mathbf{x}_i, X\rangle \tag{27}$$

6. Substitute the inner product $\langle \mathbf{x}_i, \mathbf{x}_j\rangle$ with the kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$.

$$Training: \quad \max_{\alpha,\alpha^*} -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)k(\mathbf{x}_i, \mathbf{x}_j) - \epsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n} y_i(\alpha_i - \alpha_i^*) \tag{28}$$

$$Prediction: \quad f(X) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)k(\mathbf{x}_i, X) \tag{29}$$

7. C plays a similar role as it did during classification. It is a measure of how strongly we penalize errors.

   It should be tuned for bias vs variance with model selection. The higher the value of C, the larger the tendency of SVM to penalize errors and overfit the data. The lower the value of C, the larger its tendency to ignore errors and underfit the data.

   Large C - More complex model. Low Bias, High variance.

   Small C - Less complex model. High Bias, Low Variance.

   ϵ plays the opposite role of C. The smaller the value of ϵ, the harder SVM tries to fit smaller errors around the learned SVM function, and leads to a more complex model. Smaller ϵ also leads to a less sparse solution (more support vectors).

   Small ϵ - More complex model. Low Bias, High variance.

   Large ϵ - Less complex model. High Bias, Low Variance.

Problem 5 (Boosting):

1. Gradient boosting for a loss function $L(y, F) = (y - f)^2/2$ would complete the following steps:
   (a) Initialize $\hat{f}_0(\mathbf{x}) = \arg\min_h \sum_{i=1}^{n}(y_i - h(\mathbf{x}_i))^2$
   (b) For m = 1 to M:
      i. Compute the negative gradient

$$-g_m(\mathbf{x}_i) = y_i - \hat{f}_{m-1}, i = 1 \ldots n$$

      ii. Fit a function h_m to the negative gradient by least squares

$$\hat{h}_m = \arg\min_h \sum_{i=1}^{n}((y_i - \hat{f}_{m-1}) - h(\mathbf{x}_i))^2$$

      iii. Find β to minimize the loss

$$\beta_m = \arg\min_\beta \sum_{i=1}^{n}(y_i - (\hat{f}_{m-1}(\mathbf{x}_i) + \beta\hat{h}_m(\mathbf{x}_i)))^2$$

$$\beta_m = 1$$

      iv. Update $\hat{f}$

$$\hat{f}_m(\mathbf{x}) = \hat{f}_{m-1} + \hat{h}_m(\mathbf{x})$$

(c) Output $\hat{f}_m$

2. The algorithm is iteratively fitting the residuals of the previous approximation function $y_i - f_{m-1}$.

Problem 6 (SSVM):

The Lagrange dual has $m := \sum_i |\mathcal{Y}_i|$ variables. Writing $\alpha_i(y)$ for the dual variable associated with the training example i and potential output $y \in \mathcal{Y}_i$, the dual problem is

$$\begin{aligned} &\min_{\alpha \in \mathbb{R}^m, \alpha \geq 0} f(\alpha) := \frac{\lambda}{2}||A\alpha||^2 - b^T\alpha \\ &\text{s.t.} \sum_{y \in \mathcal{Y}_i} \alpha_i(y) = 1 \ \forall i \in [n], \end{aligned} \tag{30}$$

where the matrix $A \in \mathbb{R}^{d \times m}$ consists of the m columns $A := \{\frac{1}{\lambda n}\psi_i(y) \in \mathbb{R}^d | i \in [n], y \in \mathcal{Y}_i\}$, and the vector $b \in \mathbb{R}^m$ is given by $b := (\frac{1}{n}L_i(y))_{i \in [n], y \in \mathcal{Y}_i}$.
More detailed derivation:
(1) The Lagrangian of the primal is

$$\mathcal{L}(w, \xi, \alpha) = \frac{\lambda}{2}\langle w, w \rangle + \frac{1}{n}\sum_{i=1}^{n}\xi_i + \sum_{i \in [n], y \in \mathcal{Y}_i} \frac{1}{n}\alpha_i(y)(-\xi_i + \langle w, -\psi_i(y) \rangle + \mathcal{L}_i(y)). \tag{31}$$

(2) Solve $\nabla_{(w,\xi)}\mathcal{L}(w, \xi, \alpha) = 0$. Then we get

$$\begin{aligned} \lambda w &= \sum_{i \in [n], y \in \mathcal{Y}_i} \frac{1}{n}\alpha_i(y)\psi_i(y), \\ \sum_{y \in \mathcal{Y}_i} \alpha_i(y) &= 1 \ \forall i \in [n]. \end{aligned} \tag{32}$$

(3) Plug above expression back into the Lagrangian, we obtain the Lagrange dual problem

$$\begin{aligned} &\max_{\alpha} -\frac{\lambda}{2}|| \sum_{i \in [n], y \in \mathcal{Y}_i} \alpha_i(y)\frac{\psi_i y}{\lambda n}||^2 + \sum_{i \in [n], y \in \mathcal{Y}_i} \alpha_i(y)\frac{\mathcal{L}_i(y)}{n} \\ &\text{s.t.} \sum_{y \in \mathcal{Y}_i} \alpha_i(y) = 1 \ \forall i \in [n]. \\ &\text{and } \alpha_i(y) \geq 0 \ \forall i \in [n], \forall y \in \mathcal{Y}_i, \end{aligned} \tag{33}$$

which is the same as the compact version in (30).