# ECE 517 HW 3.1

Joseph Hilland

November 2023

## 1 Assignment Details

This assignment will summarize the theoretical portion of Module 3, Support Vector Machines (SVM) for classification. The following sections from the module can be found in this summary:

- Risk and Empirical Risk
- Complexity and Overfitting
- VC Dimension
- Interpret VC Theorem
- SVM Criterion
- Analysis of Dual Solution and results
- Properties of support vectors

### 1.1 Risk and Empirical Risk

Risk in the context of machine learning is quantified with a mathematical expression that represents the expected error associated with the models predictions. The general form of the risk equation can be expressed as:

$$R(f) = \mathbb{E}[L(Y, f(X))]$$

There are typically three different sets of risks within machine learning. This includes, actual, structural and empirical.

Empirical risk can also be known as the training or sample risk. The model being trained can be measured for its average error over the dataset being used. This can be summarized as how well a model fits the training data being used. The mathematical expression for empirical risk can be expressed as:

$$R_{\text{empirical}}(f) = \frac{1}{2N} \sum_{i=1}^{N} L(y_i, f(x_i))$$

Where:

- $R_{\text{empirical}}(f)$ is the empirical risk associated with the model $f$.
- $N$ is the number of training examples.
- $(x_i, y_i)$ represents the $i$-th training example.
- $L(y_i, f(x_i))$ is the loss function that measures the discrepancy between the true label $y_i$ and the predicted value $f(x_i)$ for the $i$-th training example.

Actual risk, is the measure of the models performance on all inputs from the entire population. It can be approximated with help from the structural and empirical risk. Structural risk can be expressed as:

$$\sqrt{\frac{h(\log(2N/h)+1)-\log(\eta/4)}{N}} \tag{1}$$

Using these risk techniques we can identify how well a model will perform and if the algorithm is underfit or overfit for our data.

## 1.2 Complexity and Overfitting

Complexity refers to the sophistication of a model. Overfitting can occur where a model learns the training data too well. This ends up capturing noise or random fluctuations within the training data. This can result in a model that performs well on training data, however it will fail when given new or unseen data. This phenomenon occurs when a models algorithm becomes too complex. Keep the model as simple as possible.

## 1.3 Vapnik–Chervonenkis (VC) Dimension

The Vapnik-Chervonenkis dimension within the context of machine learning and statistical theory measures the capacity of a hypothesis class. This, is the set of all possible functions that a learning algorithm could output. The VC dimension is also related to the concept of shattering. VC dimension can also be described as the largest number of points for which the class can realize all possible binary labelings.

Shattering is said to occur when a hypothesis class "shatters" a set of points. This means for every possible labeling of those points, there exists a function in that class that can achieve that labeling.

The VC dimension can determine the max number of vectors that can be shattered by a hyperplane and gives a measure of the complexity of linear functions. If the VC Dimension of an estimator is higher than the number of vectors to be classified, then the estimator is guaranteed to overfit.

The VC dimension has the following characteristics:

- A hyperplanes VC dimension in space of dimension n is $h = n + 1$.
- VC dimesion can give a measure as to the complexity of linear functions.

- An estimator is guaranteed to overfit if the VC dimension of an estimator is higher than the number of vectors to be classified.

In summary, the VC dimension measures the capacity of a class to fit different labelings of sets of points. It assists in identifying the trade-off between model complexity and generalization performance within the context of machine learning. VC Dimension is the maximum number of vectors that can be shattered by a hyperplane.

## 1.4 VC Theorum

The Vapnik-Chervonenkis theorem provides insights into the relationships between training error and the generalization error of a machine learning model.

The theorem summarizes that given some set of m points in Rn̂, choosing any of the points as the origin, then the m points can be shattered by oriented hyperplanes if and only if the position vectors of the remaining points are linearly independent.

With the Vapnik–Chervonenkis theorem, we need to define the linear empirical risk as:

$$R_{emp}(\boldsymbol{\alpha}) = \frac{1}{2N} \sum_{n=1}^{N} \mid y - f(\mathbf{x}, \boldsymbol{\alpha}) \mid \tag{2}$$

where $f(\cdot)$ is defined so that the loss function $\mid y - f(\mathbf{x}, \boldsymbol{\alpha}) \mid$ can only take the values of 0 or 1. Then, with the probability of $1 - \eta$, the following bound holds:

$$R(\boldsymbol{\alpha}) \leq R_{emp}(\boldsymbol{\alpha}) + \sqrt{\frac{h(\log(2N/h) + 1) - \log(\eta/4)}{N}} \tag{3}$$

This is a bound on the risk with probability $1 - \eta$, so it is therefore neither guaranteed nor dependent on the probability distribution. While the left side is not computable, the right one can be easily computed provided the knowledge of $h$, where the second term of the right side is called the structural risk ($R_s$). The inductive *Principle of Structural Risk Minimization* consists then on choosing a machine whose dimension $h$ is sufficiently small, so that the bound on the risk is minimized.

## 1.5 SVM Criterion

A support vector machine constructs a hyperplane or possibly a set of hyperplanes. These hyperplanes can be used for regression or classification. The optimal hyperplane requires the usage of two other hyperplanes that are parallel from it on either side. These other two hyperplanes used for support, are located within the most extreme points between the classes.

The SVM Criterion can be expressed as:

$$min_{w,b} \frac{1}{2}|w|^2 + C \sum_{n=1}^{N} max(0, 1 - y_i(wx_i + b)) \tag{4}$$

- $\frac{1}{2}|w|^2$ represents the regularization term to encourage maximization of the margin.
- The summation term represents penalizing misclassifications. The term max will be zero when a point lies on or inside the correct margin.
- C is the regularization parameter that controls the trade-off between the size of the margin and the amount of tolerated misclassifications.

This optimization problem involves finding values of w and b that will minimize the objective function. The problem can be solved by using quadratic programming methods.

This criterion is derived from maximizing the margin between classes.

## 1.6 SVM Dual Solution, Results

In some instances data may not be linearly separable within the context of SVMs. This leads to a dual solution. The dual problem arises from the lagrangian formulation of the primal problem. The goal is to maximize the margin to certain constraints. This involves finding the lagrange multipliers that will satisfy the Karush-Kuhn-Tucker (KKT) conditions. This dual solution can be utilized when dealing with data that is not linearly separable or when kernal tricks get applied to handle complex boundaries.

The steps to forming an analysis of the dual solution are as follows:

- Primal problem: Aims to find the optimal hyperplane that separates the data with maximum margins. Typically formulated as a constrained optimization problem. This involves a cost parameter (C) for controlling the trade-off between a high-margin achievement and minimizing the wrong classification.
- Lagrangian Formulation: SVM idea is to minimize the empirical risk and structural risk through margin maximization as stated above.

$$\text{minimize } L_p(\mathbf{w}, \xi_n) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi_n \tag{5}$$

$$\text{subject to } \begin{cases} y_n\left(\mathbf{w}^\top \mathbf{x}_n + b\right) > 1 - \xi_n \\ \xi_n \geq 0 \end{cases}$$

C is a tradeoff parameter, subindex p stands for primal. Next we must then construct the functional:

$$F_Lagrange = F(w) - \boldsymbol{\alpha} g(w) \tag{6}$$

Next we need to optimize by computing the gradiant with respect to the primal variables w. And then we will null it:

$$\boldsymbol{\Delta_w} F(w) - \boldsymbol{\alpha} g(w) = 0 \tag{7}$$

This has lead to the KKT conditions. Now, we will find the value of the dual variables. The SVM primal problem is:

$$\text{minimize } L_p(\mathbf{w}, \xi_n) = \frac{1}{2}\|\mathbf{w}\|^2 + \ C\sum_{n=1}^{N}\xi_n \tag{8}$$

$$\text{subject to } \begin{cases} y_n\Big(\mathbf{w}^\top \mathbf{x}_n + b\Big) - 1 + \xi_n \geq 0 \\ \xi_n \geq 0 \end{cases}$$

However, this problem is constrained, so we must use lagrange multipliers to make it unconstrained. There are 2N constraints, and therefore we will need 2N multipliers. Namely $\alpha_n$ for the first set and $\mu_n$ for the second one. The lagrangian is now:

$$L_L(w, \epsilon_n, \alpha_n, \mu_n) = \frac{1}{2}\|\mathbf{w}\|^2 + \ C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N}\alpha_n(y_n(w^T x_n + b) - 1 + \epsilon_n) - \sum_{n=1}^{N}\mu_n\epsilon_n \tag{9}$$

This is subject to $\alpha_n, \mu_n \geq 0$ and the primal variables are w and $\epsilon_n$.

- Karush-Kuhn-Tucker Conditions: Now we must derive a solution that satisfies the KKT conditions. These conditions are necessary for optimality. The KKT conditions will include stationarity, primal feasibility, dual feasability and complementary slackness. First, lets null the gradient with respect to w:

$$\Delta_w L_L(w, \epsilon_n, \alpha_n, \mu_n) = w - \sum_{n=1}^{N}\alpha_n y_n x_n = 0 \tag{10}$$

This will end up with the result of:

$$w = \sum_{n=1}^{N}\alpha_n y_n x_n \tag{11}$$

Next we need to null the derivative with respect to the slack variables $\epsilon_n$ and b.

$$\frac{\vartheta}{\vartheta\epsilon_n}\ L_p(w, \epsilon_n, \alpha_n, \mu_n) = C - \alpha_n - \mu_n = 0 \tag{12}$$

$$\frac{d}{db}\ L_p(w, \epsilon_n, \alpha_n, \mu_n) = -\sum_{n=1}^{N}\alpha_n y_n = 0 \tag{13}$$

Now the complimentary property must be forced over the constraints:

$$\mu_n\epsilon_n = 0 \tag{14}$$

$$\alpha_n(y_n(w^T x_n + b) - 1 + \epsilon_n) = 0 \tag{15}$$

The KKT conditions are:

$$w = \sum_{n=1}^{N} \alpha_n y_n x_n \tag{16}$$

$$C - \alpha_n - \mu_n = 0 \tag{17}$$

$$\sum_{n=1}^{N} \alpha_n y_n = 0 \tag{18}$$

$$\mu_n \epsilon_n = 0 \tag{19}$$

$$alpha_n(y_n(w^T x_n + b) - 1 + \epsilon_n) = 0 \tag{20}$$

$$\alpha_n \geq 0, \mu_n \geq 0, \epsilon \geq 0 \tag{21}$$

- The dual solution: From 16 and 18 KKT conditions above:

$$C - \alpha_n - \mu_n = 0 \tag{22}$$

$$\mu_n \epsilon_n = 0 \tag{23}$$

we can see that if $\epsilon_n > 0$ then $\alpha_n = C$. With 19, if a sample is on the margin then $0 < \alpha_n < C$. And if the sample has been classified well then $\epsilon_n = 0$, and 19 determines that $\alpha_n = 0$. Finally, if we plug 15 into the Lagrangian, we can solve for the dual solution:

$$L_d = -\frac{1}{2} \sum_{n=1}^{N} \sum_{n'=1}^{N} y_n \alpha_n x_n^T x_{n'} \alpha_{n'} y_{n'} + \sum_{n=1}^{N} \alpha_n \tag{24}$$

with the constraint $\alpha_n \geq 0$.

This analysis has shown that we can construct a Lagrange functional from the primal expression of the SVM functional. We can then find the support vectors and dual expression of the classifier as a function of them by computing the derivative of the Lagrangian with respect to the primal parameters.

## 1.7 SV Properties

Support vectors are crucial for determining the optimal hyperplane. This hyperplane separates different classes in the SVM. The support vectors are the points that will lie closest to the margin. If a datapoint is well within the boundary (support hyperplane), the penalizing factor $\xi_n$ is 0. Otherwise, if the datapoint is on the other side, this factor $\xi_n$ is equal to its distance between the datapoint and the support hyperplane, i.e., $\xi_n \geq 0$ and $\alpha_n = C$. If a sample is on the margin, $0 < \alpha_n = C$. Finally, if a sample is outside the margin, $\xi_n = 0$ and $\alpha_n = 0$.

## 1.8 Conclusion

SVMs are powerful machine learning algorithms. They are used for classification and regression tasks. SVMs aim to find a hyperplane that will best separate data into classes. SVMs also aim to maximize the margin between classes.