CSC411: Assignment 3

Due on Sunday, December 3rd, 2017

Student Name: **Gokul K. Kaushik**

Student Number: **999878191**

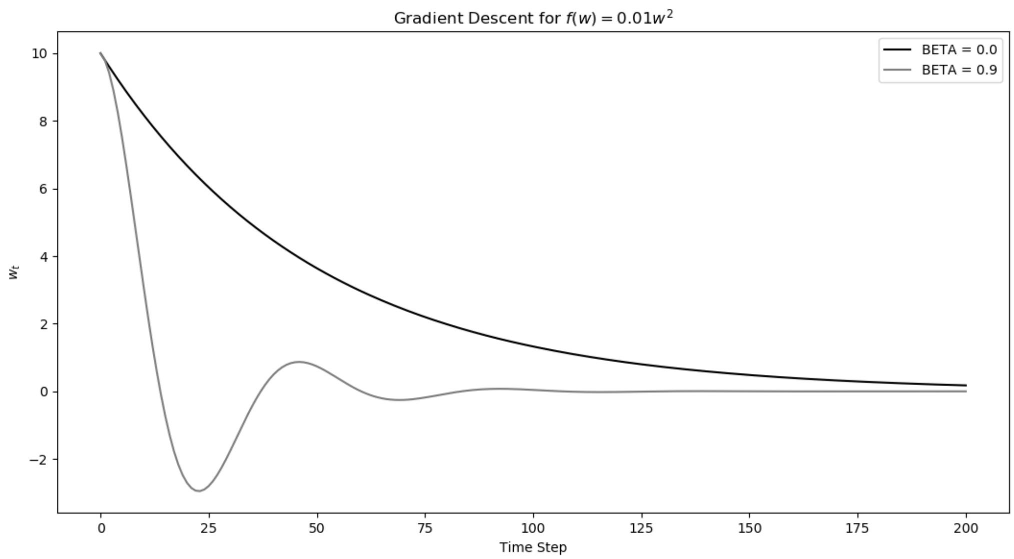
Table of Contents

1 - 20 Newsgroup Predictions

2 - Training SVM with SGD

2.1 - SGD with Momentum

Plotting w_t for each time step t by applying the iterative stochastic gradient-descent on $f(w) = 0.01w^2$, we get the following graph (for $\beta = 0.1$ and $\beta = 0.9$ for upto 200 time-steps):



2.2 -Training SVM

2.3 - Apply 4-vs-9 Digits on MNIST

Two SVM models were trained using gradient descent. Both models had the following properties:

1. A learning rate of $\alpha = 0.05$
2. A penalty of $C = 1.0$
3. Mini-batch sizes of $m = 100$ and $T = 500$

Model 1 and 2 had different β values.

1. **Model 1**: $\beta = 0$
2. **Model 2**: $\beta = 0.1$

For the first model use = 0 and for the second use = 0.1. For both of the trained models report the following:

2.3.1 - Training Loss

1. **Model 1** ($\beta = 0$) **Training Loss**: 0.3829
2. **Model 2** ($\beta = 0.1$) **Training Loss**: 0.3743

2.3.2 - Test Loss

1. **Model 1** ($\beta = 0$) **Test Loss**: 0.3748
2. **Model 2** ($\beta = 0.1$) **Test Loss**: 0.3672

2.3.3 - Classification Accuracy on the Training Set

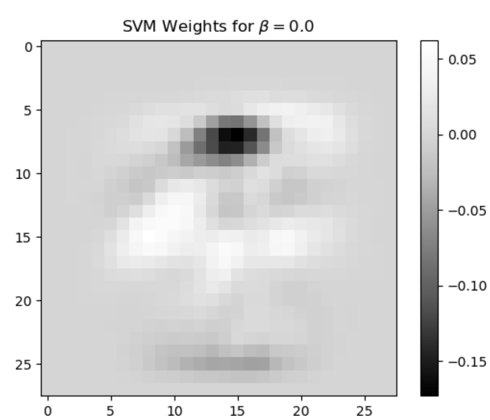
1. **Model 1** ($\beta = 0$) **Classification Accuracy on Training Set**: 0.9139
2. **Model 2** ($\beta = 0.1$) **Classification Accuracy on Training Set**: 0.9127

2.3.4 - Classification Accruacy on the Test Set

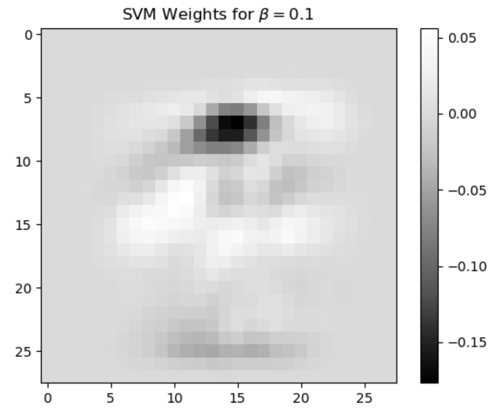
1. **Model 1** ($\beta = 0$) **Classification Accuracy on Test Set**: 0.9154
2. **Model 2** ($\beta = 0.1$) **Classification Accuracy on Test Set**: 0.9125

2.3.5 - Plot w as a 28 $\times$ 28 image

$\mathbf{w}$ for Model 1 ($\beta = 0$):



$\mathbf{w}$ for Model 2 ($\beta = 0.1$):



3 - Kernels

3.1 - Positive Semi definite and Quadratic Form

Prove that a symmetric matrix $K \in \mathbb{R}^{d \times d}$ is a positive semi definite iff for all vectors x we have $\mathbf{x}^T K \mathbf{x} \geq 0$.

Proof:

$$K\mathbf{x} = \lambda \mathbf{x}$$

where λ is the eigenvalue and $\mathbf{x}$ is the eigenvector.

Suppose $\mathbf{x}$ is an eigenvector of K and replacing $K\mathbf{v}$ with $\lambda \mathbf{v}$ (from the definition of an eigenvector and eigenvalue above):

$$\mathbf{x}^T K \mathbf{x} = \mathbf{x}^T \mathbf{x} \lambda$$

$$\mathbf{x}^T K \mathbf{x} = |\mathbf{x}|^2 \lambda$$

Therefore, as $|\mathbf{x}|^2 \geq 0$, for the equation $\mathbf{x}^T K \mathbf{x} \geq 0$, the eigenvalue must be: $\lambda \geq 0$. Since it is a semi-definite matrix, where the eigenvalues ≥ 0, this holds true.

3.2 - Kernel Properties

3.2.1 - Prove Property $k(\mathbf{x}, \mathbf{y}) = \alpha$ is a kernel for $\alpha > 0$

$$\phi(\mathbf{x}) = \sqrt{\alpha}$$

$$k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$$

$$k(\mathbf{x}, \mathbf{y}) = \sqrt{\alpha}\sqrt{\alpha}$$

$$k(\mathbf{x}, \mathbf{y}) = \alpha$$

3.2.2 - Prove Property $k(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) \cdot f(\mathbf{y})$ is a kernel for $f : \mathbb{R}^d \rightarrow \mathbb{R}$

$$\phi(\mathbf{x}) = f(\mathbf{x})$$

$$k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$$

$$k(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) \cdot f(\mathbf{y})$$

3.2.3 - Prove Property If $k_1(\mathbf{x}, \mathbf{y})$ and $k_2(\mathbf{x}, \mathbf{y})$ are kernels then $k(\mathbf{x}, \mathbf{y}) = a \cdot k_1((\mathbf{x}, \mathbf{y}) + b \cdot k_2((\mathbf{x}, \mathbf{y})$ for $a, b > 0$ is a kernel

Let $\mathbf{K}_1$ and $\mathbf{K}_2$ be gram matrices

Therefore:

$$\mathbf{x}^T \mathbf{K}_1 \mathbf{x} \geq 0$$

and

$$\mathbf{x}^T \mathbf{K}_2 \mathbf{x} \geq 0$$

Let:

$$\mathbf{K}_3(\mathbf{x}, \mathbf{y}) = a\mathbf{K}_1(\mathbf{x}, \mathbf{y}) + b\mathbf{K}_2(\mathbf{x}, \mathbf{y})$$

$$\mathbf{K}_3 = a\mathbf{K}_1 + b\mathbf{K}_2$$

$$\mathbf{x}^T \mathbf{K}_3 \mathbf{x} = \mathbf{x}^T (a\mathbf{K}_1 + b\mathbf{K}_2)\mathbf{x}$$

$$\mathbf{x}^T \mathbf{K}_3 \mathbf{x} = a\mathbf{x}^T \mathbf{K}_1 \mathbf{x} + b\mathbf{x}^T \mathbf{K}_2 \mathbf{x}$$

and using 3.2.1's proof, $\mathbf{K}_3(\mathbf{x}, \mathbf{y})$ is a kernel, we get:

$$a\mathbf{x}^T \mathbf{K}_1 \mathbf{x} + b\mathbf{x}^T \mathbf{K}_2 \mathbf{x} \geq 0 \text{ for } a, b > 0$$

3.2.4 - Prove Property If $k_1(\mathbf{x}, \mathbf{y})$ is a kernel then $k(\mathbf{x}, \mathbf{y}) = \frac{k_1(\mathbf{x}, \mathbf{y})}{\sqrt{k_1(\mathbf{x}, \mathbf{x})}\sqrt{k_1(\mathbf{y}, \mathbf{y})}}$ is a kernel

Let:

$$k_1(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$$

$$\langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle = \phi(\mathbf{x}) \cdot \phi(\mathbf{y})$$

In equation $k_1(\mathbf{x}, \mathbf{y})$, replacing all the instances of $k_1(\mathbf{x}, \mathbf{y})$ with $\phi(\mathbf{x}) \cdot \phi(\mathbf{y})$:

$$k(\mathbf{x}, \mathbf{y}) = \frac{k_1(\mathbf{x}, \mathbf{y})}{\sqrt{k_1(\mathbf{x}, \mathbf{x})}\sqrt{k_1(\mathbf{y}, \mathbf{y})}}$$

$$k(\mathbf{x}, \mathbf{y}) = \frac{\phi(\mathbf{x}) \cdot \phi(\mathbf{y})}{\sqrt{\phi(\mathbf{x}) \cdot \phi(\mathbf{x})}\sqrt{\phi(\mathbf{y}) \cdot \phi(\mathbf{y})}}$$

$$k(\mathbf{x}, \mathbf{y}) = \frac{\phi(\mathbf{x}) \cdot \phi(\mathbf{y})}{|\phi(\mathbf{x})||\phi(\mathbf{y})|}$$

$$k(\mathbf{x}, \mathbf{y}) = \frac{k_1(\mathbf{x}, \mathbf{y})}{|\phi(\mathbf{x})||\phi(\mathbf{y})|}$$

Let $\mathbf{K}_1$ for matrix $k_1(\mathbf{x}, \mathbf{y})$ be a gram matrix. Therefore:

$$\mathbf{x}^T \mathbf{K}_1 \mathbf{x} \geq 0$$

Multiplying each entry of k_1 will result in a positive value. Therefore, $k(\mathbf{x}, \mathbf{y})$ is a kernel.