Machine Learning Worksheet 9

Support Vector Machines

1 Optimisation and convexity

In the lecture you learned the concept of a convex function. By tighening the condition we obtain the concept of a strictly convex function.

A function $f : \mathcal{X} \to \mathbb{R}$ defined on a convex set $\mathcal{X}$ is called *strictly convex* if, for any two points $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$ with $\boldsymbol{x} \neq \boldsymbol{y}$ and any $t \in]0, 1[$

$$f(t\boldsymbol{x} + (1-t)\boldsymbol{y}) < tf(\boldsymbol{x}) + (1-t)f(\boldsymbol{y})\,.$$

It can be shown that for differentiable functions this is equivalent to the condition that for every $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$ with $\boldsymbol{x} \neq \boldsymbol{y}$ we have

$$f(\boldsymbol{y}) > f(\boldsymbol{x}) + (\boldsymbol{y} - \boldsymbol{x})^T \nabla f(\boldsymbol{x})\,.$$

Problem 1: Consider the graph of a strictly convex function. How can you define strict convexity geometrically? How does it compare to (non-strict) convexity?

Problem 2: If a function $f(\boldsymbol{x})$ is convex how many points $\boldsymbol{x}^*$ at most can satisfy the equation $f(\boldsymbol{x}^*) = \min_{\boldsymbol{x}} f(\boldsymbol{x})$? Proof your answer.

Problem 3: If a function $f(\boldsymbol{x})$ is strictly convex how many points $\boldsymbol{x}^*$ at most can satisfy the equation $f(\boldsymbol{x}^*) = \min_{\boldsymbol{x}} f(\boldsymbol{x})$? Proof your answer.

Problem 4: Give examples of a non-strictly convex function and a strictly convex function. Provide definitions and graphs of your functions.

2 The Gaussian kernel for classification

Problem 5: Let us define a feature transformation $\boldsymbol{\phi_n} : \mathbb{R} \to \mathbb{R}^n$ as follows:

$$\boldsymbol{\phi_n}(x) = \left\{ e^{-x^2/2\sigma^2}, e^{-x^2/2\sigma^2}\frac{x}{\sigma}, \frac{e^{-x^2/2\sigma^2}\left(\frac{x}{\sigma}\right)^2}{\sqrt{2}}, \ldots, \frac{e^{-x^2/2\sigma^2}\left(\frac{x}{\sigma}\right)^i}{\sqrt{i!}}, \ldots, \frac{e^{-x^2/2\sigma^2}\left(\frac{x}{\sigma}\right)^n}{\sqrt{n!}} \right\}$$

Suppose we let $n \to \infty$ and define a new feature transformation:

$$\boldsymbol{\phi_\infty}(x) = \left\{ e^{-x^2/2\sigma^2}, e^{-x^2/2\sigma^2}\frac{x}{\sigma}, \frac{e^{-x^2/2\sigma^2}\left(\frac{x}{\sigma}\right)^2}{\sqrt{2}}, \ldots, \frac{e^{-x^2/2\sigma^2}\left(\frac{x}{\sigma}\right)^i}{\sqrt{i!}}, \ldots \right\}$$

Can we directly apply this feature transformation to data? Explain why or why not!

Problem 6: From the lecture, we know that we can express a linear classifier using only inner products of input vectors in the transformed feature space. It would be great if we could somehow use the feature space obtained by the feature transformation $\boldsymbol{\phi_\infty}$. However, to do this, we must be able to compute the inner product of samples in this infinite feature space. We define the inner product between two *infinite* vectors $\boldsymbol{\phi_\infty}(x)$ and $\boldsymbol{\phi_\infty}(y)$ as the infinite sum given in the following equation:

$$K(x, y) = \sum_{i=1}^{\infty} \boldsymbol{\phi_{\infty,i}}(x)\, \boldsymbol{\phi_{\infty,i}}(y)$$

What is the explicit form of $K(x, y)$? (Hint: Think of the Taylor series of e^x.) With such a high dimensional feature space, should we be concerned about overfitting?

Problem 7: Can any *finite* set of points be linearly separated in the feature space defined by $\boldsymbol{\phi_\infty}$ if σ can be chosen freely?

3 The SVM optimisation problem

Problem 8: Consider the SVM optimisation problem with slack variables (1-norm soft-margin) from the lecture. Show that the duality gap is zero for this problem. Why is this important? What would happen if this was not the case?

4 A concrete Support Vector Machine (tutor exercise)

Now you are given a data set with data from a single feature x in $\mathbb{R}$ and corresponding labels $y \in \{+1, -1\}$. Data points for $+1$ are at $-3, -2, 3$ and data points for -1 are at $-1, 0, 1$.

Problem 9: Can this data set in its current feature space be separated using a linear separator? Why/why not?

Let's define a simple feature map $\boldsymbol{\phi}(x) = (x, x^2)$ that transforms points in $\mathbb{R}$ to points in $\mathbb{R}^2$.

Problem 10: After applying $\boldsymbol{\phi}$ to the data, can it now be separated using a linear separator? Why/why not? (Plotting the data may help you with your answer.)

Problem 11: *Construct* a maximum-margin separating hyperplane (i.e. you do not need to solve a quadratic program). Clearly mark the support vectors. Also draw the resulting decision boundary in the original feature space. Is it possible to add another point to the training set in such a way, that the hyperplane *does not* change? Why/why not?

Problem 12: For this specific training set write down the SVM optimisation problem, the Lagrangian, the Lagrange dual function and the dual problem. Do not introduce slack variables.

Problem 13: Write down the KKT conditions for this training set explicitly and verify that the maximum-margin hyperplane you constructed satisfies them.