

Machine Learning Worksheet 9

Support Vector Machines

1 Optimization and convexity

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

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

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

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

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

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 $f(\mathbf{x})$ is convex how many points \mathbf{x}^* at most can satisfy the equation $f(\mathbf{x}^*) = \min_{\mathbf{x}} f(\mathbf{x})$? Proof your answer.

Problem 3: If a function $f(\mathbf{x})$ is strictly convex how many points \mathbf{x}^* at most can satisfy the equation $f(\mathbf{x}^*) = \min_{\mathbf{x}} f(\mathbf{x})$? 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 $\phi_n : \mathbb{R} \rightarrow \mathbb{R}^n$ as follows:

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

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

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

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 ϕ_∞ . 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 $\phi_\infty(x)$ and $\phi_\infty(y)$ as the infinite sum given in the following equation:

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

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 ϕ_∞ 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?