Exercises
Introduction to Machine Learning
SS 2019

Series 3, Mar 25th, 2019 (SVM, Kernels)

Institute for Machine Learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Andreas Krause
Web: `https://las.inf.ethz.ch/teaching/introml-s19`
Email questions to:
max.paulus@inf.ethz.ch, joanna.ficek@inf.ethz.ch

Note: These are sample solutions. If you solved the problem in a different way it doesn't necessarily mean that your solution is wrong.

Problem 1 (SVM):

Consider the surrogate loss

$$l_s(\mathbf{w}; \mathbf{x}, y) = \begin{cases} 0, & \text{for } \mathrm{sign}(\mathbf{w}^T\mathbf{x}) = y \\ \sqrt{-y\mathbf{w}^T\mathbf{x}}, & \text{for } \mathrm{sign}(\mathbf{w}^T\mathbf{x}) \neq y \end{cases}$$

a) **Is l_s convex?**
To check whether l_s is convex, we can look at $f(x) = \sqrt{x}$.
A way to show that $f(x) = \sqrt{x}$ is not convex is to show that $-f(x)$ is convex.

$$\begin{aligned} \sqrt{tx_1 + (1-t)x_2} &> t\sqrt{x_1} + (1-t)\sqrt{x_2} \\ tx_1 + (1-t)x_2 &> t^2x_1 + (1-t)^2x_2 + t2(1-t)\sqrt{x_1x2} \\ x_1 + x_2 &> 2\sqrt{x_1x_2} \\ (\sqrt{x_1} - \sqrt{x_2})^2 &> 0 \end{aligned}$$

Hence, $f(x) = \sqrt{x}$ is concave and so is l_s.

Is l_s differentiable?
Let's differentiate with respect to $y\mathbf{w}^T\mathbf{x}$. If $sign(\mathbf{w}^T\mathbf{x}) = y$, $l_s'(\mathbf{w}; \mathbf{x}, y) = 0$.

If $sign(\mathbf{w}^T\mathbf{x}) \neq y$, $l_s'(\mathbf{w}; \mathbf{x}, y) = \frac{1}{2}(-y\mathbf{w}^T\mathbf{x})^{-\frac{1}{2}}(-1) = -\frac{1}{2\sqrt{-y\mathbf{w}^T\mathbf{x}}}$.

To check differentiability we need to check the limit at point 0.
Let $z = y\mathbf{w}^T\mathbf{x}$. Then, $\lim_{z\to 0^-} -\dfrac{1}{\sqrt{z}} = -\infty$. Hence, l_s is not differentiable at $y\mathbf{w}^T\mathbf{x} = 0$.

b) **Derive $\nabla l_s(w, x, y)$.**
Although l_s not differentiable at $y\mathbf{w}^T\mathbf{x} = 0$, the subgradient exists and hence (stochastic) gradient descent converges. To derive the subgradient let's rewrite the function l_s as $l_s(\mathbf{w}; \mathbf{x}, y) = max(0, \sqrt{-y\mathbf{w}^T\mathbf{x}})$. Now let $f(z) = max(0, \sqrt{-yz})$ and $g(\mathbf{w}) = \mathbf{w}^T\mathbf{x}$. We use the chain rule

$$\frac{\partial}{\partial w_i} f(g(\mathbf{w})) = \frac{\partial f}{\partial z}\frac{\partial g}{\partial w_i} \ .$$

We get

$$\frac{\partial f}{\partial z} = \begin{cases} 0, & \text{for } \mathrm{sign}(z) = y \\ -\frac{y}{2\sqrt{-yz}}, & \text{for } \mathrm{sign}(z) \neq y \end{cases}$$

and $\frac{\partial g}{\partial w_i} = x_i$. Hence,

$$\frac{\partial f(g(\mathbf{w}))}{\partial w_i} = \begin{cases} 0, & \text{for } \mathrm{sign}(\mathbf{w}^T\mathbf{x}) = y \\ -\frac{y\mathbf{x}}{2\sqrt{-y\mathbf{w}^T\mathbf{x}}}, & \text{for } \mathrm{sign}(\mathbf{w}^T\mathbf{x}) \neq y \end{cases} \ .$$

c) The exercise suggests to train an SVM, where we penalise the margin violation given by $(1 - y\mathbf{w}^T\mathbf{x})_+ = \max(1 - y\mathbf{w}^T\mathbf{x}, 0)$ not linearly but with the square root instead. Correspondingly, our modified SVM seeks to optimise the following objective

$$L(w) = \frac{1}{n}\sum_{i=1}^{n}\sqrt{(1 - y\mathbf{w}^T\mathbf{x})_+} + \lambda\|w\|^2 \tag{1}$$

We could try to optimise this objective using stochastic gradient descent, for example

Initialize (e.g. $\mathbf{w}_1 = 0$)
For $t = 1, 2, \ldots$ do

Pick $i_t \sim Unif(1, \ldots, n)$
if $y_{i_t}\mathbf{w_t}^T\mathbf{x}_{i_t} < 1$
$\mathbf{w}_{t+1} = \mathbf{w}_t(1 - \eta_t 2\lambda\mathbf{w}_t) + \eta_t\frac{y_i\mathbf{x_i}}{2\sqrt{(1-y_i\mathbf{w}^T\mathbf{x_i})}}$
else
$\mathbf{w}_{t+1} = \mathbf{w}_t(1 - \eta_t 2\lambda\mathbf{w}_t)$

Why may this modification not be a good idea? You can see that the weight update due to margin violations gets rescaled as a result of the modification by the factor $\frac{1}{2\sqrt{(1-y_i\mathbf{w}^T\mathbf{x_i})}}$. This factor is small when the margin violation is large and large when the margin violation is small, which may make training this modified SVM troublesome.

Problem 2 (Kernels):

a) Since each polynomial term is a product of kernels with positive coefficients, the proof follows from the rules of addition and multiplication yielding valid kernels (see Tutorial V).

b) We can use the Taylor expansion around 0:

$$\begin{aligned} exp(k(x,y)) &= exp(0) + exp(0)k(x,y) + \frac{exp(0)}{2!}(k(x,y))^2 + \ldots \\ &= 1 + k(x,y) + \frac{1}{2}(k(x,y))^2 + \frac{1}{6}(k(x,y))^3 + \ldots \end{aligned}$$

An exponential of a kernel is an infinite series of additions and multiplications of that kernel and hence, is a valid kernel (follows from the rules of addition and multiplication yielding valid kernels, see Tutorial V).

c) Since $k(x,y)$ is a valid kernel, we can define a feature map $\phi(.)$, such that $k(x,y) = \langle\phi(x), \phi(y)\rangle$.
Now,

$$k_c(x,y) = f(x)k(x,y)f(y) = f(y)f(x)\langle\phi(x), \phi(y)\rangle = f(y)\langle f(x)\phi(x), \phi(y)\rangle = \langle f(x)\phi(x), f(y)\phi(y)\rangle \ .$$

Hence, with the new feature map $\phi_c(.) = f(.)\phi(.)$, $k_c(x,y)$ is a valid kernel (symmetry and positive definiteness properties didn't change).

d) We know that $k(x,y)$ is a valid kernel and hence, on any set of vectors (also transformed ones) it yields a valid kernel.

Problem 3 (Kernels: Past Exam):

1. Let $\mathbf{x}^T = (x_1 \ldots, x_d)$, and note that,

$$(\mathbf{x}^T\mathbf{x}' + 1)^2 = (\sum_i x_i x_i' + 1)^2 = 1 + 2\sum_i x_i x_i' + \sum_i \sum_j (x_i x_j) \cdot (x_i' x_j')$$
$$= 1 + \sum_i (\sqrt{2}x_i) \cdot (\sqrt{2}x_i') + \sum_i \sum_j (x_i x_j) \cdot (x_i' x_j')$$

Thus $\phi(x)$ can be a vector of dimension $1 + d + d^2$ such that its first entry is 1, its next d entries are $\sqrt{2}x_i$, and its remaining d^2 entries are $x_i x_j$.

2. First, we get $\phi(\mathbf{x})$ for each $\mathbf{x}$.

   (a) $\phi([-3, 4]) = (-3, 4, 5)$
   (b) $\phi([1, 0]) = (1, 0, 1)$

   Now we get the inner products;

   (a) $\phi([-3, 4])^T \phi([-3, 4]) = 50$
   (b) $\phi([-3, 4])^T \phi([1, 0]) = 2$
   (c) $\phi([1, 0])^T \phi([1, 0]) = 2$

   And now the Gram matrix Φ is simply given by $\Phi_{i,j} = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$; using the above:

$$\begin{pmatrix} 50 & 2 \\ 2 & 2 \end{pmatrix}$$