Theoretical task 1

All solutions should be short, mathematically precise and contain proof unless qualitative explanation/intuition is needed.

1. Suppose $x \in \mathbb{R}^D$ is a feature vector. Prove that *whitening transformation* $f = \Sigma^{-1/2}(x - \mu)$, where $\mu = \mathbb{E}x$, $\Sigma = cov[x, x]$, will give new feature vector f with properties:

   (a) $\mathbb{E}f = \mathbf{0}$ (all zeroes vector)

   (b) $cov[f, f] = I$ (identity matrix)

2. Consider training set $x_1, ...x_N$ and some linear subspace L_K with lower dimensionality $K \leq D$. Let $x_i = p_i + h_i$ where p_i are projections of x_i onto L_K and h_i are orthogonal complements. Suppose we perform optimization over different K-dimensional subspaces L_K. Prove equivalence of the following two optimization tasks:

   (a) $\sum_{i=1}^{N} \|h_i\|^2 \to \min_{L_K}$

   (b) $\sum_{i=1}^{N} \|p_i\|^2 \to \max_{L_K}$

   Comment: $\|z\|$ is L_2 norm of vector z.

3. Write stochastic gradient descent with minibatch size=1 for the following losses:

   (a) $\mathcal{L}(M) = e^{-M}$

   (b) $\mathcal{L}(M) = [1 - M]_+$

   Why classification quality (evaluated by by margin) on object from the minibatch cannot decrease?

4. Prove that $K(x, x') = e^{-\gamma\langle x-x', x-x'\rangle}$, $\gamma > 0$ is a Mercer kernel.

   Hint: use operations generating new kernels out of existing kernels.

5. Consider a binary classifier $\widehat{y}(x) = sign(g(x) - \mu)$ with discriminant function $g(x)$ and some threshold μ. Suppose you know $TPR(\mu)$ and $FPR(\mu)$. Now consider an inverted classifier $\tilde{y}(x) = sign(\mu - g(x))$. Write out $TPR(\mu)$ and $FPR(\mu)$ measures for it in terms of original classifier. Explain.

6. Consider multiclass classification performed by M classifiers $f_1(x), ...f_M(x)$. Let probability of mistake be constant $p \in (0, \frac{1}{2})$: $p(f_m(x) \neq y) = p\,\forall m$ and suppose all models make mistakes or correct guesses independently of each other. Let $F(x)$ be majority voting aggregation function (voting for most popular class among predicted by $f_1(x), ...f_M(x)$). Prove that $\forall (x, y)\, p(F(x) \neq y) \to 0$ as $M \to \infty$.

   Hint: use central limit theorem, consider fraction of errors. What can be said about its expectation and variance?