ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Pattern Recognition and Machine Learning

Homework 1 (Week 2)

Lecturer: Matthias Seeger
Assistants: Abdallah Elguindy, Mathias Humbert, Young Jun Ko, Xinchao Wang

Solutions have to be handed in in week 4, **at the start of the exercise session, Thursday October 10, 1pm, INJ 218**. Late hand-ins will not be considered for grading and score zero points. Hand in solutions on paper (neat handwriting or typed). Graded solutions are returned in the exercise session in the week following the hand-in or, in exceptional circumstances, two weeks after hand-in.
Homework problems must be solved independently, teamwork is not permitted. Do not copy or let others copy from you. According to EPFL regulations, copying is considered fraud and leads to failure of the course (for all students involved).

Exercise 1

In this exercise, you will prove an important matrix identity. If m is much smaller than n, this identity can be used to very efficiently invert matrices of a special structure. It is most frequently used for $m = 1$.
Suppose that $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times m}$ are matrices, where $m < n$.

1. [**3 points**] Suppose that the matrix $\boldsymbol{I} + \boldsymbol{A}\boldsymbol{B}^T \in \mathbb{R}^{n\times n}$ is invertible. Here, $\boldsymbol{I}$ is the identity matrix. Prove the following identity:
$$\left(\boldsymbol{I} + \boldsymbol{A}\boldsymbol{B}^T\right)^{-1} = \boldsymbol{I} - \boldsymbol{A}\left(\boldsymbol{I} + \boldsymbol{B}^T\boldsymbol{A}\right)^{-1}\boldsymbol{B}^T.$$

2. [**2 points**] Now suppose that $\boldsymbol{C} \in \mathbb{R}^{n\times n}$ is invertible, and $\boldsymbol{C} + \boldsymbol{A}\boldsymbol{B}^T$ is invertible as well. Prove the following generalization of the previous identity:
$$\left(\boldsymbol{C} + \boldsymbol{A}\boldsymbol{B}^T\right)^{-1} = \boldsymbol{C}^{-1} - \boldsymbol{C}^{-1}\boldsymbol{A}\left(\boldsymbol{I} + \boldsymbol{B}^T\boldsymbol{C}^{-1}\boldsymbol{A}\right)^{-1}\boldsymbol{B}^T\boldsymbol{C}^{-1}.$$
Hint: Recall that $(\boldsymbol{C}\boldsymbol{X})^{-1} = \boldsymbol{X}^{-1}\boldsymbol{C}^{-1}$, and use the identity you have established in the first part.

Exercise 2

The Winnow algorithm is an alternative to the perceptron algorithm for learning linear classifiers with positive weights. It makes use of multiplicative rather than additive updates, which can lead to faster convergence on sparse data.
Input points are $\boldsymbol{x}_i \in \{0, 1\}^d$, so each feature is either 0 or 1. In practice, d can be very large, but typical points $\boldsymbol{x}_i$ are sparse (most entries are 0). Targets $t_i \in \{-1, +1\}$ are binary. The classifier is
$$f(\boldsymbol{x}) = \operatorname{sgn} y(\boldsymbol{x}), \quad y(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b, \quad w_j > 0,\ j = 1, \dots, d.$$
In the simple version of Winnow discussed here, all weights w_j are positive, and the bias parameter b is not learned, but fixed up front to $b = -d$, where d is the dimensionality of the feature space.

Suppose we are given data $\{(\boldsymbol{x}_i, t_i) \mid i = 1, \ldots, n\}$. Under the assumption that there is some solution $(\boldsymbol{w}_*, b)$ with $w_j^* > 0$ and $t_i(\boldsymbol{w}_*^T \boldsymbol{x}_i + b) > 0$ for all $i = 1, \ldots, n$, then the following algorithm finds a solution after finitely many updates. Winnow cycles over patterns in some ordering. Suppose the current weight vector is $\boldsymbol{w}$ and we visit $(\boldsymbol{x}_i, t_i)$. Let $y_i = \boldsymbol{w}^T \boldsymbol{x}_i + b$.

- If $t_i y_i > 0$ (example classified correctly): Do nothing.
- If $y_i \leq 0$ and $t_i = +1$ (error, y_i too small): Multiply by 2 those components w_j of $\boldsymbol{w}$ for which $x_{ij} = 1$, leave the others ($x_{ij} = 0$) unchanged.

$$w_j \leftarrow w_j 2^{x_{ij}}, \quad j = 1, \ldots, d.$$

- If $y_i \geq 0$ and $t_i = -1$ (error, y_i too large): Divide by 2 those components w_j of $\boldsymbol{w}$ for which $x_{ij} = 1$, leave the others ($x_{ij} = 0$) unchanged.

$$w_j \leftarrow w_j 2^{-x_{ij}}, \quad j = 1, \ldots, d.$$

The update rule can be summarized as

$$w_j \leftarrow w_j 2^{t_i x_{ij}}, \quad \text{if} \;\; t_i y_i \leq 0.$$

1. **[4 points]** For a simple email SPAM filter, you decide to employ five features of a text message: $\boldsymbol{x} \in \{0,1\}^5$ (so $d = 5$). Here, $x_j = 1$ iff the j-th word in { and, viagra, the, of, nigeria } is present. You collect a dataset of $n = 6$ emails.

i	t_i	and	viagra	the	of	nigeria
1	+1	1	1	0	1	1
2	-1	0	0	1	1	0
3	+1	0	1	1	0	0
4	-1	1	0	0	1	0
5	+1	1	0	1	0	1
6	-1	1	0	1	1	0

Run the Winnow algorithm until convergence, visiting patterns in the order $1, 2, 3, 4, 5, 6$. For each $(\boldsymbol{x}_i, t_i)$, state whether there is an update, and if so, what the new weight vector $\boldsymbol{w}$ is. Start from $\boldsymbol{w} = [1, 1, 1, 1, 1]^T$ and use $b = -5$ (this bias parameter is not changed).
Note: You may have to run over the data more than once.

Exercise 3

A set $\mathcal{A} \subset \mathbb{R}^d$ is *convex* if for any points $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{A}$, we have that

$$\lambda \boldsymbol{x} + (1 - \lambda)\boldsymbol{y} \in \mathcal{A} \quad \forall \lambda \in [0, 1]$$

For example, a halfspace

$$\mathcal{H} = \left\{ \boldsymbol{x} \in \mathbb{R}^d \mid \boldsymbol{w}^T \boldsymbol{x} + b \geq 0 \right\}$$

is a convex set.

1. **[2 points]** Suppose that $\mathcal{H}_1$, $\mathcal{H}_2$ (given by $(\boldsymbol{w}_1, b_1)$, $(\boldsymbol{w}_2, b_2)$ respectively) are arbitrary halfspaces in $\mathbb{R}^d$. For each of the following sets, either prove its convexity or show that it is not convex by giving a counterexample (for the latter, a drawing in $\mathbb{R}^2$ suffices)

   a) Intersection $\mathcal{H}_1 \cap \mathcal{H}_2$

   b) Union $\mathcal{H}_1 \cup \mathcal{H}_2$

   c) Complement of union

$$(\mathcal{H}_1 \cup \mathcal{H}_2)^c = \{\boldsymbol{x} \mid \boldsymbol{x} \notin \mathcal{H}_1 \cup \mathcal{H}_2\}$$

2. **[2 points]** Suppose that

$$\mathcal{A} = \left\{[x, y, z]^T \mid z > 0\right\}$$

is a convex set, where the last component z is positive for all members. Consider the mapping $P([x, y, z]^T) := [x/z, y/z]^T$ from $\mathcal{A}$ to $\mathbb{R}^2$. Prove that for any two points $[x_1, y_1, z_1]^T, [x_2, y_2, z_2]^T \in \mathcal{A}$, the line segment

$$\lambda_1[x_1, y_1, z_1]^T + (1 - \lambda_1)[x_2, y_2, z_2]^T, \quad \lambda_1 \in [0, 1]$$

is mapped one-to-one onto the line segment

$$\lambda_2[x_1/z_1, y_1/z_1]^T + (1 - \lambda_2)[x_2/z_2, y_2/z_2]^T, \quad \lambda_2 \in [0, 1].$$

Hint: Determine λ_2 for λ_1. The function $\lambda_1 \mapsto \lambda_2$ is strictly increasing (you do not have to show that). Argue that it maps $[0, 1]$ onto $[0, 1]$.

3. **[2 points]** Prove that for a convex set $\mathcal{A}$ as defined in the previous part, the set

$$P(\mathcal{A}) = \left\{[x/z, y/z]^T \mid [x, y, z]^T \in \mathcal{A}\right\}$$

is convex as well. Use what you have shown above.