1. Stochastic gradient descent, when used with the hinge loss, leads to which update rule?
   Options:
   • Perceptron
   • Widrow’s Adaline
   • Winnow
   • AdaGrad
   Ans: Perceptron

2. Let’s assume that we are using the standard Averaged Perceptron algorithm for training and testing (prediction). The training data consists of m examples. Let’s further assume that it makes k mistakes on the training data. Now, how many weight vectors do we require to predict the label for a test instance?
   Options:
   • $O(k)$
   • $O(1)$
   • $O(k^2)$
   • $O(m)$
   • Not enough information
   Ans: $O(k)$

3. Which of the following properties is true about the (original) Perceptron algorithm?
   Options:
   • If the data given to the Perceptron is linearly separable the Perceptron will stop making mistakes after some number of examples.
   • The Perceptron always converges to the best linear separator for a given dataset.
   • The convergence criteria for Perceptron depends on the initial value of the weight vector.
   • If the dataset is not linearly separable, the Perceptron algorithm learns the linear separator with least misclassifications.
   Ans: If the data given to the Perceptron is linearly separable the Perceptron will stop making mistakes after some number of examples.
4. In a mistake-driven algorithm, if we make a mistake on example $x_i$ with label $y_i$, we update the weights $w$ so that we now do not make a mistake on this example if we see it again.

**Options:**

- False
- True

**Ans:** False

5. Let the learned weights after the perceptron algorithm finishes training be weight vector $w$. Suppose that the bias term is 0. If we scale $w$ by a positive constant factor (multiply each element of $w$ with $c$), then the new set of weights

**Options:**

- produces the exact same classification for all the data points
- may output different classification results for some data points

**Ans:** produces the exact same classification for all the data points

6. Let the learned weights after the perceptron algorithm finishes training be weight vector $w$. Suppose that the bias term is 0. If we translate $w$ by a positive constant factor $c$ (add $c$ to each element of $w$), then the new set of weights

**Options:**

- may output different classification results for some data points
- produces the exact same classification for all the data points

**Ans:** may output different classification results for some data points