Supervised Learning, COMP0078, Main Exam, 20/21

Answer ALL SIX questions. You may use results from the notes without reproving them, however, add a citation.

**Notation:** Let $[m] := \{1, \ldots, m\}$. We also overload notation so that

$$[\texttt{pred}] := \begin{cases} 1 & \texttt{pred} \text{ is true} \\ 0 & \texttt{pred} \text{ is false} \end{cases}.$$

Marks for each part of each question are indicated in square brackets.

1. We consider problem of empirical risk (error) minimisation for multivariate simple polynomial functions. For all $\boldsymbol{\alpha} \in \mathbb{R}^n$ define $h_{\boldsymbol{\alpha}} : (0, \infty)^n \rightarrow (0, \infty)$ as

$$h_{\boldsymbol{\alpha}}(\boldsymbol{x}) = x_1^{\alpha_1} x_2^{\alpha_2} \times \cdots \times x_n^{\alpha_n},$$

the hypothesis space,

$$\mathcal{H}_{\text{exp}} := \{h_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \mathbb{R}^n\}$$

and the error function $\ell(y, \hat{y}) := \log(\frac{y}{\hat{y}})^2$. Given a dataset $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m) \subset (0, \infty)^n \times (0, \infty)$ design an efficient (polynomial-time) algorithm to perform empirical risk minimisation with respect to $\mathcal{H}_{\text{exp}}$ and $\ell$. Argue that your algorithm is correct.

[10 marks]

2. a. Suppose $K : [0, \mathbb{R}) \times [0, \mathbb{R}) \rightarrow \mathbb{R}$ is a kernel.

   i. Is $K(x^3, t^3)$ a kernel? Explain why or why not.

   ii. Is $K(\sqrt{x}, \sqrt{t})$ a kernel? Explain why or why not. Note for the purpose of this question $\sqrt{\cdot}$ is single-valued and returns the positive root.

[5 marks]

b. A training set $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)$ is *generic* iff $\boldsymbol{x}_i = \boldsymbol{x}_j \Rightarrow y_i = y_j$. Consider the following kernel function $K_a(\boldsymbol{x}, \boldsymbol{t}) := \prod_{i=1}^{n}(1 + (x_i t_i) + (1 - x_i)(1 - t_i))$.

   i. Argue that $K_a : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ is a kernel.

   ii. Given a generic training set $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m) \in \{0, 1\}^n \times \{0, 1\}$ does there necessarily exist a vector $\boldsymbol{\alpha} \in \mathbb{R}^m$ such that

$$\sum_{j=1}^{m} \left( \sum_{i=1}^{m} \alpha_i K_a(\boldsymbol{x}_i, \boldsymbol{x}_j) - y_j \right)^2 = 0\,?$$

   Provide an argument to justify your answer.

[5 marks]

3. *Linear support vector machines* (SVMS)

Assume that the set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m \subset \mathbb{R}^2 \times \{-1, 1\}$ of binary examples is strictly linearly separable by a line going through the origin, that is, there exists a vector $\boldsymbol{w} \in \mathbb{R}^2$ such that the linear function $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$, $\boldsymbol{x} \in \mathbb{R}^2$ has the property that $y_i f(\boldsymbol{x}_i) > 0$ for every $i = 1, \ldots, m$. Consider the optimisation problem (linearly separable SVM):

$$\mathbf{P1}: \quad \text{minimise} \left\{ \frac{1}{2} \boldsymbol{w}^\top \boldsymbol{w} \; : \; y_i \boldsymbol{w}^\top \boldsymbol{x}_i \geq 1, i = 1, \ldots, m \right\}.$$

a. Argue that the above problem has a unique solution. Describe the geometric meaning of this solution.

[5 marks]

b. Show that the vector $\boldsymbol{w}$ solving problem **P1** has the form $\boldsymbol{w} = \sum_{i=1}^m c_i y_i \boldsymbol{x}_i$ where $c_1, \ldots, c_m$ are some nonnegative coefficients. [HINT: use the method of Lagrange multipliers]

[5 marks]

c. Show that the coefficients $c_1, \ldots, c_m$ in the above formula solve the optimization problem

$$\mathbf{P2}: \; \max \left\{ -\frac{1}{2} \sum_{i,j=1}^m c_i c_j y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j + \sum_{i=1}^m c_i \; : \; c_j \geq 0, \; j = 1, \ldots, m \right\}.$$

Finally, if $(\hat{c}_1, \ldots, \hat{c}_m)$ is the solution to this problem and $\hat{\boldsymbol{w}}$ is the solution to problem P1, argue that $\hat{\boldsymbol{w}}^\top \hat{\boldsymbol{w}} = \sum_{i=1}^m \hat{c}_i$.

[5 marks]

[Question 3 cont. over page]

d. Now drop the linear separability assumption. Consider the following data dependent representation: $\boldsymbol{z}_i = (\boldsymbol{x}_i, \rho e_i) \in \mathbb{R}^{d+m}$, where $\rho$ is a positive parameter and $e_i$ is the $m$ dimensional vector all of whose components are equal to zero except for the $i$-th component which is equal to $1$. Argue that the binary data $\{(\boldsymbol{z}_i, y_i)\}_{i=1}^m$ are linearly separable. Furthermore show that the corresponding problem,

$$\mathbf{P3}: \quad \min_{\boldsymbol{v}\in\mathbb{R}^{2+m}} \left\{\frac{1}{2}\boldsymbol{v}^\top \boldsymbol{v} \;:\; y_i \boldsymbol{v}^\top \boldsymbol{z}_i \geq 1, i = 1, \ldots, m\right\}$$

is equivalent to the problem

$$\mathbf{P4}: \quad \min_{\boldsymbol{w}\in\mathbb{R}^2, \xi\in\mathbb{R}^m} \left\{\frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} + \frac{1}{2\rho^2}\sum_{i=1}^m \xi_i^2 \;:\; y_i \boldsymbol{w}^\top \boldsymbol{x}_i \geq 1 - \xi_i, i = 1, \ldots, m\right\}.$$

Then provide an interpretation of the latter problem as regularisation problem, indicating which is the loss function used. (2-5 sentences recommended.)

[5 marks]

4. *Adaboost.*

Recall the following notation,

- $\alpha_t$ : the weight on weak learner $t$ where $\alpha_t \in \mathbb{R}$.
- $D_t(i)$ : the weight on example $i$ at time $t$ where $\sum_{i=1}^{m} D_t(i) = 1$
- $\epsilon_t$ : "weighted error of weak learner $h_t(\cdot)$ at time $t$"

$$\epsilon_t := \sum_{i=1}^{m} D_t(i) \left[h_t(\boldsymbol{x}_i) \neq y_i\right]$$

- training error at time $T$,

$$\frac{1}{m} \sum_{i=1}^{m} [H(\boldsymbol{x}_i) \neq y_i]$$

i. Briefly discuss the role of the "$\alpha$" weights and "$D$" weights in Adaboost (2-5 sentences recommended). Your discussion should remark on the significance of relatively high or low weight values.

[5 marks]

ii. Let $\epsilon^* = \max_{t \in [T]} \epsilon_t$, use $\epsilon^*$ to bound the training error at time $T$.

[5 marks]

iii. *Look-ahead weighted error.*

$$\hat{\epsilon}_t := \sum_{i=1}^{m} D_{t+1}(i) \left[h_t(\boldsymbol{x}_i) \neq y_i\right],$$

What, if anything, can we infer about the $\hat{\epsilon}_t$ when $t = T$. Explain your reasoning.

[10 marks]

5. The problem of designing efficient algorithms that obtain (expected) non-trivial regret bounds for linear classification with respect to the "0-1" loss (misclassifications) seems to be a difficult problem. In this question we consider simpler problems.

   We have the following definitions the 0-1 loss is $\ell_{01}(y, \hat{y}) := [y \neq \hat{y}]$, the hypothesis class of 2-norm bounded linear *classifiers* over a set $\mathcal{X} \subset \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| = 1\}$,

$$\widehat{\mathcal{H}}_{\mathcal{X},U} = \{h_{\boldsymbol{u}} : \boldsymbol{u} \in \mathbb{R}^n, \|\boldsymbol{u}\| \leq U, \forall \boldsymbol{x} \in \mathcal{X} : |\boldsymbol{u} \cdot \boldsymbol{x}| \geq 1\}$$

   where $h_{\boldsymbol{u}} : \mathcal{X} \to \mathbb{R}$ and $h_{\boldsymbol{u}}(\boldsymbol{x}) := \operatorname{sign}(\boldsymbol{u} \cdot \boldsymbol{x})$, and the corresponding hypothesis class of linear *interpolants*

$$\bar{\mathcal{H}}_{\mathcal{X},U} = \{h_{\boldsymbol{u}} : \boldsymbol{u} \in \mathbb{R}^n, \|\boldsymbol{u}\| \leq U, \forall \boldsymbol{x} \in \mathcal{X} : |\boldsymbol{u} \cdot \boldsymbol{x}| = 1\},$$

   intuitively the functions in $\widehat{\mathcal{H}}$ classify the points in $\mathcal{X}$ with a margin $\geq 1$ and the functions in $\bar{\mathcal{H}}$ interpolate the points in $\mathcal{X}$ to be exactly 1 or -1. Also observe that $\bar{\mathcal{H}}_{\mathcal{X},U} \subseteq \widehat{\mathcal{H}}_{\mathcal{X},U}$.

   a. i. Discuss why giving regret bounds for predicting linear classifiers with respect to the 0-1 loss may be more difficult than than with the hinge loss. Please limit your discussion to 1-2 sentences.

      ii. Discuss why from a "user's perspective" it may be more useful to have a 0-1 loss than a hinge loss bound. Please limit your discussion to 1-3 sentences.

[5 marks]

[Question 5 cont. on next page]

b. *Simplifying by restricting the hypothesis class.*

**Protocol:**

**Nature** selects $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T \in \mathcal{X}$ and $y_1, y_2, \ldots, y_T \in \{-1, 1\}$.

For $t = 1$ To $T$ Do

| | |
|---|---|
| Receive pattern | $\boldsymbol{x}_t \in \mathcal{X}$ |
| Predict | $\hat{y}_t \in \{-1, 1\}$ |
| Receive label | $y_t \in \{-1, 1\}$ |

Design a polynomial-time randomised algorithm with a good upper bound on the expected regret,

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_{01}(y_t, \hat{y}_t)\right] - \min_{h \in \bar{\mathcal{H}}_{\mathcal{X},U}} \sum_{t=1}^{T} \ell_{01}(y_t, h(\boldsymbol{x}_t))$$

give your algorithm design and argument for the upper bound.

[5 marks]

[Question 5 cont. over page]

[Question 5 cont.]

c. *Simplifying by loosening the definition of regret.* In this part we generalise regret to allow a leading term $c(U)$ in front of the loss of the best hypothesis this leading term can only depend on $U$.

**Protocol:**

**Nature** selects $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T \in \mathcal{X}$ and $y_1, y_2, \ldots, y_T \in \{-1, 1\}$.

For $t = 1$ To $T$ Do

| | |
|---|---|
| Receive pattern | $\boldsymbol{x}_t \in \mathcal{X}$ |
| Predict | $\hat{y}_t \in \{-1, 1\}$ |
| Receive label | $y_t \in \{-1, 1\}$ |

Design a polynomial-time algorithm with a good upper bound on the $c(U)$-regret,

$$\sum_{t=1}^{T} \ell_{01}(y_t, \hat{y}_t) - c(U) \min_{h \in \widehat{\mathcal{H}}_{\mathcal{X},U}} \sum_{t=1}^{T} \ell_{01}(y_t, h(\boldsymbol{x}_t)) \, .$$

Observe that the introduction of $c(U)$ generalises the notion of regret where $c(U) = 1$ recovers the usual notion of regret. Thus the aim is that $c(U)$ is small while still guaranteeing that,

$$\frac{\sum_{t=1}^{T} \ell_{01}(y_t, \hat{y}_t)}{T} \leq c(U) \min_{h \in \widehat{\mathcal{H}}_{\mathcal{X},U}} \frac{\sum_{t=1}^{T} \ell_{01}(y_t, h(\boldsymbol{x}_t))}{T} \text{ as } T \to \infty \, .$$

Give your algorithm design and argument for the upper bound.

[10 marks]

6. a. Given an example of a hypothesis class $\mathcal{H}$ with $|\mathcal{H}| = \infty$ and $\text{vcdim}(\mathcal{H}) = 1$. Argue that it has $\text{vcdim}(\mathcal{H}) = 1$.

[5 marks]

b. Suppose we have hypothesis class $\mathcal{H}$ with $|\mathcal{H}| = \infty$ and $\text{vcdim}(\mathcal{H}) = 1$. What does does this imply, if anything, about the possible existence of an online algorithm with a mistake bound for this hypothesis class. Explain your reasoning.

[5 marks]

c. Suppose we have hypothesis class $\mathcal{H}$ and an algorithm with a mistake bound of $B$ for $\mathcal{H}$. What, if anything, does this imply about the pac-learnability of the hypothesis class. Explain your reasoning.

[10 marks]