TC1 - Exam 2020-2021 - 2 hours

Caio Corro

All documents are allowed. No phone, no computer. You **must** use formal notation and you **must** defines all mathematical terms you use.

1 Course questions (20 points)

1. **(1 point)** Why do we want the training problem to be a convex or concave optimization problem?
2. **(2 points)** What important property of convex functions can impact the optimization algorithm? Why?
3. **(1 point)** What does "support vector" refer to in Support Vector Machines?
4. **(2 point)** Why are slack variable introduced in the constrained formulation of SVM ? (you can use a picture to illustrate your explanation)
5. **(1 point)** In the course and lab exercise, we assumed that there was a feature always equal to one so we don't have to explicitly introduce a bias term in our scoring function. However we could have an explicit bias term instead, i.e. a scoring function of the form $s(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{a} + b$. Give a reason why we shouldn't apply regularization to the bias term when training a SVM (hint: think about the geometric interpretation of the binary SVM classifier).
6. **(2 points)** Why is it interesting to optimize the dual formulation of a SVM via the coordinate ascent algorithm? (at least 2 reasons)
7. **(2 points)** Let $\boldsymbol{w} \in \mathbb{R}^k$ be a vector of scores and $\hat{\boldsymbol{y}} \in \mathbb{R}^k$ be a one-hot vector representing the gold output. The negative log-likelihood loss for a softmax prediction function can be written as:

   $$l(\hat{\boldsymbol{y}}, \boldsymbol{w}) = -\log\left(\hat{\boldsymbol{y}}^\top \text{softmax}(\boldsymbol{w})\right)$$

   Compute the gradient $\nabla_{\boldsymbol{w}} l(\hat{\boldsymbol{y}}, \boldsymbol{w})$. How can it be interpreted?
8. **(3 points)** We now consider the hinge loss:

   $$h(\hat{\boldsymbol{y}}, \boldsymbol{w}) = \max(0, -\hat{\boldsymbol{y}}^\top \boldsymbol{w} + m + \max_{\boldsymbol{y} \in \mathcal{Y}(k) \setminus \{\hat{\boldsymbol{y}}\}} \boldsymbol{y}^\top \boldsymbol{w})$$

   What is (one) sub-gradient $\nabla_{\boldsymbol{w}} h(\hat{y}, \boldsymbol{w})$? (We use the gradient notation just the simplify notation) How can it be interpreted? Compare with the gradient $\nabla_{\boldsymbol{w}} l(\hat{y}, \boldsymbol{w})$ of the previous function.
9. **(3 points)** Computing the hinge loss function requires to solve an optimization problem over the set $\boldsymbol{y} \in \mathcal{Y}(k) \setminus \{\hat{\boldsymbol{y}}\}$. Propose and justify a definition of the hinge loss where this optimization is over $\boldsymbol{y} \in \mathcal{Y}(k)$, i.e. propose a formulation of the hinge loss of the form:

   $$h(\hat{\boldsymbol{y}}, \boldsymbol{w}) = \max(0, ... \max_{\boldsymbol{y} \in \mathcal{Y}(k)} ...)$$

10. **(1 point)** Why we shouldn't use the negative log-likelihood loss function when training a model based on the sparsemax prediction function?
11. **(2 points)** Consider these three different loss functions for structured prediction: negative log-likelihood, SVM and SVM with rescaled margin. In which setting could you use each of these loss functions? Why?

2 Convex analysis (14 points)

1. **(2 points)** What are two different interpretations of the subgradient of a convex function?
2. **(2 points)** Let $f(x) = -\log x$ be the negative logarithm function. Compute the Fenchel conjugate of the negative logarithm function. The Fenchel conjugate is defined as follows:

   $$f^*(y) = \sup_x xy - f(x)$$

3. **(2 points)** Compute the Fenchel conjugate of the Fenchel conjugate of the negative logarithm function (i.e. the biconjugate of $f(x)$). Is this result expected?

4. **(1 points)** If we consider the function $h(x) = \log x$ instead, would we expect to have $h^{**}(x) = h(x)$? Why? (no need to compute anything for this question)

5. **(2 points)** Explain how we can rely on Fenchel conjugates to define prediction functions (and prediction functions only) and why this is useful.

6. **(3 points)** Let $f : X \to \mathbb{R}$ be a convex function where $X \subset \mathbb{R}^n$ is a convex subset of $\mathbb{R}^n$. Prove that its extended-value extension $\tilde{f} : \mathbb{R}^n \to \mathbb{R}$ is convex.

$$\tilde{f}(x) = \begin{cases} f(x) & \text{if } x \in X \\ \infty & \text{otherwise} \end{cases}$$

7. **(2 points)** Why is it useful to rely on extended-value extension of functions? Why is it important that $\tilde{f}(x)$ is convex? (you can give an example)

3 Fenchel-Young losses (16 points)

Note: You are expected to describe both the regularized prediction function q_Ω and the associated Fenchel-Young loss L_Ω in both questions.

1. **(8 points)** We assume the following scoring and prediction function for regression:

$$\begin{aligned} s(\boldsymbol{x}) &= \boldsymbol{a}^\top \boldsymbol{x} & s &: \mathbb{R}^n \to \mathbb{R} \\ q(w) &= w & s &: \mathbb{R} \to \mathbb{R} \end{aligned}$$

   where $\boldsymbol{a} \in \mathbb{R}^n$ are the parameters of the model. The standard training loss is the squared loss defined as $l(\hat{y}, y) = \frac{1}{2}\|\hat{y} - y\|_2^2$ where $\hat{y}$ is the gold output and y is the predicted output. Prove formally that we can express this prediction and loss function in the Fenchel-Young loss framework.

2. **(8 points)** We assume the following scoring and prediction for probabilistic binary classification:

$$\begin{aligned} s(\boldsymbol{x}) &= \boldsymbol{a}^\top \boldsymbol{x} & s &: \mathbb{R}^n \to \mathbb{R} \\ q(w) &= \sigma(w) & s &: \mathbb{R} \to]0, 1[\end{aligned}$$

   where $\sigma(w) = \frac{\exp(w)}{1+\exp(w)}$ is the sigmoid function and $\boldsymbol{a} \in \mathbb{R}^n$ are the parameters of the model. In other words, the function q returns the parameter μ of a Bernoulli distribution defined as follows:

$$p(y; \mu) = \mu^y (1 - \mu)^{(1-y)}$$

   The entropy of a Bernoulli distribution is defined as follows:

$$H[p(\cdot; \mu)] = -y \log y - (1 - y) \log(1 - y)$$

   Build formally the regularized prediction function and the Fenchel-Young loss associated with this probabilistic binary classifier.

Cheat sheet: KKT conditions

Assume we have a maximization problem defined as follows:

$$\begin{aligned} \max_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & g^{(i)}(\boldsymbol{x}) \geq 0 & \forall 1 \leq i \leq m \\ & h^{(i)}(\boldsymbol{x}) = 0 & \forall 1 \leq i \leq n \end{aligned}$$

where $\boldsymbol{x} \in \mathbb{R}^k$, $g^{(i)} \geq 0$ is a set of m inequality and $h^{(i)}(\boldsymbol{x}) = 0$ is a set of n equality. An optimal solution $\boldsymbol{x}^* \in \mathbb{R}^k$ of the mathematical program satisfies the following constraints:

$$\begin{aligned} &\text{(stationarity)} && \forall i: \quad \frac{\partial}{\partial x_i} f(x^*) + \sum_j \mu_j \frac{\partial}{\partial x_i} g^{(j)}(\boldsymbol{x}^*) - \sum_j \lambda_j \frac{\partial}{\partial x_i} h^{(j)}(\boldsymbol{x}^*) = 0 \\ &\text{(primal feasibility)} && \forall i: \quad g^{(i)}(\boldsymbol{x}) \geq 0 \\ &&& \forall i: \quad h^{(i)}(\boldsymbol{x}) = 0 \\ &\text{(dual feasibility)} && \forall i: \quad \mu_i \geq 0 \\ &\text{(complementary slackness)} && \sum_i \mu_i g^{(i)}(\boldsymbol{x}^*) \geq 0 \end{aligned}$$

where $\mu \in \mathbb{R}_+^m$ and $\lambda \in \mathbb{R}^n$ are dual variables associated with primal inequalities and equalities, respectively.