TC1 - Exam 2021-2022 - 2 hours and 30 minutes

Caio Corro

No document, no phone, no computer. You **must** use formal notation and you **must** define all mathematical terms you use.

1 Course questions (11 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 variables introduced in the constrained formulation of SVM ? (you can use a picture to illustrate your explanation - for this question you are allowed to not be very formal, but the answer must be clear)
5. **(2 points)** Why is it interesting to optimize the dual formulation of a SVM via the coordinate ascent algorithm? (at least 2 reasons)
6. **(2 points)** Let $\boldsymbol{w} \in \mathbb{R}^k$ be a vector of scores and $\hat{\boldsymbol{y}} \in E(k)$ be a one-hot vector representing the gold output. The negative log-likelihood loss for the softmax prediction function can be written as:

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

   Compute the gradient $\nabla_{\boldsymbol{w}} l(\boldsymbol{w}; \hat{\boldsymbol{y}})$. How can it be interpreted?
7. **(1 point)** What could go wrong if we use the negative log-likelihood loss function when using sparsemax as a prediction function?

2 Convex analysis (24 points)

1. **(2 points)** Let $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be the negative logarithm function, i.e. $f(x) = -\log x$. Compute (explicitly) the Fenchel conjugate and the biconjugate of f. Is this result expected?
2. **(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)
3. **(2 points)** Let $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be a function defined as $f(x) = ax + b$ where $a \in \mathbb{R}$ and $b \in \mathbb{R}$. Compute (explicitly) the Fenchel conjuguate and biconjugate of f.
4. **(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 real-valued extension $\tilde{f} : \mathbb{R}^n \to \mathbb{R}$ is convex.

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

5. **(2 points)** Why is it useful to rely on extended real-valued extension of functions? Why is it important that $\tilde{f}(x)$ is convex? (you can give an example)
6. **(2 points)** Let $f(x) = |x|$. Prove that g defined as follows is a subgradient of f at x:

   $$g = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \\ 0 & \text{otherwise} \end{cases}$$

7. **(2 points)** Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function defined as follows:

$$f(\boldsymbol{x}) = \sum_i f_i(x_i)$$

where $f_1, ..., f_n$ are convex functions defined as $f_i : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$. Prove that the conjugate of f can be written as:

$$f^*(\boldsymbol{y}) = \sum_i f_i^*(y_i)$$

8. **(1 points)** In which situation is the result of the previous function useful in Machine Learning? (one example, from the course)

9. **(2 points)** Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function defined as follows:

$$f(\boldsymbol{x}) = \alpha h(\boldsymbol{x}) \qquad \text{with} \qquad \alpha > 0$$

Prove that $f^*(\boldsymbol{y}) = \alpha h^*(\boldsymbol{y}/\alpha)$.

10. **(1 points)** In which situation is the result of the previous function useful in Machine Learning? (one example, from the course)

11. **(2 points)** Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function. Let $\boldsymbol{y} \in \operatorname{dom} f^*$. Prove that $\boldsymbol{g} = \arg\max_{\boldsymbol{x} \in dom f} \boldsymbol{y}^\top \boldsymbol{x} - f(\boldsymbol{x})$ is a subgradient of $f^*(\boldsymbol{y})$.

12. **(1 points)** In which situation is the result of the previous function useful in Machine Learning? (one example, from the course)

13. **(2 points)** Let $f : \mathbb{R} \to \mathbb{R}$ be the function defined as follows:

$$f(w) = \max(0, 1 - w)$$

Compute the Fenchel conjugate of f.

14. **(1 points)** In which situation is the result of the previous function useful in Machine Learning? (one example, from the course)

3 Fenchel-Young losses (24 points)

Note: for questions 3 and 4, you are expected to describe both the regularized prediction function $\hat{y}_\Omega$ and the associated Fenchel-Young loss L_Ω.

1. **(4 points)** Let $\Omega : \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ be a convex regularizer with $\operatorname{dom} \Omega \subseteq \triangle^k$. Prove that $\hat{y}_\Omega$ is invariant to constants, that is:

$$\forall w \in \mathbb{R}^k, c \in \mathbb{R} : \quad \hat{y}_\Omega(\boldsymbol{w}) = \hat{y}_\Omega(\boldsymbol{w} + c)$$

2. Let $\Omega : \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ be a convex regularizer. Prove that the associated Fenchel-Young loss is always non-negative, that is:

$$\forall \boldsymbol{w} \in \mathbb{R}^k, \boldsymbol{y} \in E(k) : \quad L_\Omega(\boldsymbol{w}; \boldsymbol{y}) \geq 0$$

3. **(4 points) (8 points)** We assume the following scoring and prediction function for regression:

$$\hat{w}(\boldsymbol{x}) = \boldsymbol{a}^\top \boldsymbol{x} \qquad \hat{w} : \mathbb{R}^n \to \mathbb{R}$$
$$\hat{y}(w) = w \qquad \hat{y} : \mathbb{R} \to \mathbb{R}$$

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

4. **(8 points)** We assume the following scoring and prediction for probabilistic binary classification:

$$\hat{w}(\boldsymbol{x}) = \boldsymbol{a}^\top \boldsymbol{x} \qquad \hat{w} : \mathbb{R}^n \to \mathbb{R}$$
$$\hat{y}(w) = \sigma(w) \qquad \hat{y} : \mathbb{R} \to]0, 1[$$

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 $\hat{y}(w)$ 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 and derive their closed form expressions (as you would implement them).

Cheat sheet

4 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.

5 Fenchel conjugate

Let $f : \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ be a function. The the Fenchel conjugate of f is the function $f^* : \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ defined as:

$$f^*(\boldsymbol{y}) = \sup_{\boldsymbol{x} \in \operatorname{dom} f} \boldsymbol{x}^\top \boldsymbol{y} - f(\boldsymbol{x})$$

6 Subgradient

Let $f : \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ be a function. A vector $\boldsymbol{g} \in \mathbb{R}^k$ is a subgradient of f at $\boldsymbol{x} \in \mathbb{R}^k$ if and only if:

$$\forall \boldsymbol{x}' \in \mathbb{R}^k: \quad f(\boldsymbol{x}') \geq f(\boldsymbol{x}) + \boldsymbol{g}^\top (\boldsymbol{x}' - \boldsymbol{x})$$

7 Regularized prediction functions and Fenchel-Young losses

Let $\Omega : \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ be a convex regularizer. The prediction function associated with Ω is:

$$\hat{\boldsymbol{y}}(\boldsymbol{w}) = \arg\max_{\boldsymbol{\mu} \in \operatorname{dom} \Omega} \boldsymbol{w}^\top \boldsymbol{\mu} - \Omega(\boldsymbol{\mu})$$

Then Fenchel-Young loss associated with Ω is:

$$L_\Omega(\boldsymbol{w}; \boldsymbol{y}) = \Omega^*(\boldsymbol{w}) + \Omega(\boldsymbol{y}) - \boldsymbol{y}^\top \boldsymbol{w}$$

8 Inverse function

Let σ be the sigmoid funciton, i.e. $\sigma(w) = \frac{\exp(w)}{1+\exp(w)}$. The inverse of the sigmoid is: $\sigma^{-1}(\mu) = \log \frac{\mu}{1-\mu}$.