Introduction à l'apprentissage automatique - TD2

Caio Corro

1 Gradient computation (part 2)

In this exercise, we focus no multiclass classification with k classes. Losses for these problems are defined as functions of the form $\ell : E(k) \times \mathbb{R}^k \to \mathbb{R}_+$.

1. The negative log-likelihood for multiclass classification is defined as $\ell_{nll}(\boldsymbol{y}, \boldsymbol{w}) = -\boldsymbol{y}^\top \boldsymbol{w} + \log \sum_j \exp(w_j)$. In practice, we have $\boldsymbol{w} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$. Compute gradient of the loss function wrt to $\boldsymbol{A}$ and $\boldsymbol{b}$ (note: as $\boldsymbol{A}$ is a matrix, gradient is an abuse of terminology).

2. The hinge loss for multiclass classification is defines as:
$$\ell_{hinge}(\boldsymbol{y}, \boldsymbol{w}) = \max(0, -\boldsymbol{w}^\top \boldsymbol{y} + m + \max_{\boldsymbol{y}' \in E(k) \setminus \{\boldsymbol{y}\}} \boldsymbol{w}^\top \boldsymbol{y}')$$
where $m \in \mathbb{R}_+$ is the margin. We assume $m = 1$. Prove that the following formulation is equivalent:
$$\ell_{aug}(\boldsymbol{y}, \boldsymbol{w}) = -\boldsymbol{w}^\top \boldsymbol{y} + \max_{\boldsymbol{y}' \in E(k)} (\boldsymbol{w} + (\boldsymbol{1} - \boldsymbol{y}))^\top \boldsymbol{y}'$$
we call the maximization problem in this loss "loss-augmented inference".

3. The hinge-loss function is not differentiable. However, we saw in the course that we can "approximate" the gradient of a function defined as a maximum of differentiable functions. More precisely, let $f(\boldsymbol{u}) = \max(f_1(\boldsymbol{u}), f_2(\boldsymbol{u}), ..., f_n(\boldsymbol{u}))$ be a convex function were each f_i is convex and differentiable. Let $\mathrm{I}(\boldsymbol{u}) = \{i \in \{1..n\} | f_i(\boldsymbol{u}) = f(\boldsymbol{u})$ the set of function indices reaching the maximum value for input $\boldsymbol{u}$. Then, $\forall i \in \mathrm{I}(\boldsymbol{u}) : \nabla f_i(\boldsymbol{u}) \in \partial f(\boldsymbol{u})$, i.e. the gradient of any function $f_i(\boldsymbol{u})$ s.t. $i \in \mathrm{I}(\boldsymbol{u})$ is a subgradient of $f(\boldsymbol{u})$. Using this fact, compute a subgradient of the hinge loss for multiclass classification wrt to parameters $\boldsymbol{A}$ and $\boldsymbol{b}$.
$$\ell_{aug}(\boldsymbol{y}, \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$$
This subgradient will be used instead of the gradient for learning. We abuse notation and note these subgradient as gradients.

2 Linear regression

To study linear regression, it is often useful to represent data in a matrix-vector notation. Let $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ be a vector containing the input feature of datapoints, one datapoint per row, and $\boldsymbol{y} \in \mathbb{R}^n$ be the associated output values. We assume $\boldsymbol{X}$ and $\boldsymbol{y}$ are our training data. If we ignore the intercept term (or add it as a feature), learning an unregularized linear regression model reduces to solving the following optimization problem:

$$\boldsymbol{a}^* = \arg\min_{\boldsymbol{a} \in \mathbb{R}^d} \frac{1}{2} \|\boldsymbol{X}\boldsymbol{a} - \boldsymbol{y}\|_2^2$$

1. Use first-order optimality condition to compute a closed form solution for the training problem. Under what conditions is this approach feasible? Under what conditions is this approach computationally challenging?

2. A matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ is invertible if and only if its columns are linearly independent, i.e. there exist no $\boldsymbol{v} \in \mathbb{R}^d$, $\boldsymbol{v} \neq 0$, such that $\boldsymbol{X}\boldsymbol{v} = 0$. Show that if $\boldsymbol{X}$ is not invertible, then $\boldsymbol{X}^\top \boldsymbol{X}$ is not invertible.

3. We denote $\boldsymbol{X}_i$ the i-th row of matrix i. We assume each row consists of the value 1 followed by a $d-1$ one-hot vector, i.e. the one-hot encoding of a $d-1$ categorical feature, i.e. the first column of $\boldsymbol{X}$ correspond to an implicit bias term. Prove that the column of $\boldsymbol{X}$ are not linearly independent. What can we deduce?

4. We now assume the following regularized training problem:

$$\boldsymbol{a}^* = \operatorname*{arg\,min}_{\boldsymbol{a} \in \mathbb{R}^d} \|\boldsymbol{X}\boldsymbol{a} - \boldsymbol{y}\|_2^2 + \beta \sum_{i=2}^{d} \boldsymbol{a}_i^2$$

   Warning: note that we don't regularize the weight in $\boldsymbol{a}$ associated with the implicit bias feature and we removed the $\frac{1}{2}$ term to simplify computation.

   Prove that, with this additional regularization term, the problem now has a closed form solution. **Hint:** you need to rewrite the problem so it "looks like" an unregularized problem.

5. Let D be a training dataset and consider the following training problem:

$$\operatorname*{arg\,min}_{\boldsymbol{a} \in \mathbb{R}^d, b} \frac{1}{|D|} \sum_{(\boldsymbol{x}, y) \in D} \frac{1}{2}(y - (\langle \boldsymbol{a}, \boldsymbol{x} \rangle + b))^2$$

   Derive the closed form solution for updates in the coordinate descent algorithm.

6. Derive the coordinate descent updates with additional L2 regularization in the objective.