Statistical Machine Learning — Prof. Ulrike von Luxburg + Hein group

Exercise Sheet 6 - 5.6.2024 — due: 12.6.2024

Exercise 1 Logistic Regression

We consider a problem of a (binary) logistic regression. Given data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \{-1, 1\}$ we have a linear classifier f with classification rule $f(x) = \operatorname{sign}(\langle w, x \rangle)$. The vector w is obtained as a solution to the convex optimization problem

$$\min_w L(w) = \min_w \sum_{i=1}^{n} \ln(1 + e^{-y_i \langle w, x_i \rangle}). \tag{1}$$

However, the decision boundary is linear (in the data points). Suppose that we want a more complex classifier; thus, we introduce a feature mapping $\phi(x) : \mathbb{R}^d \to \mathbb{R}^m$. The optimization problem will still be convex in w:

$$\min_w \sum_{i=1}^{n} \ln(1 + e^{-y_i \langle w, \phi(x_i) \rangle}). \tag{2}$$

Then the classification rule will be $f(x) = \operatorname{sign}(\langle w, \phi(x) \rangle)$ and is no longer linear in x in general. The mapping ϕ should satisfy that the log ratios of conditional probabilities are linear in $\phi(x)$ for the logistic regression to work well. The mappings ϕ that you will derive later are not unique.

- (**3 Points**) Show that objective L in (1) is convex in w by deriving the Hessian matrix and showing that it is positive semi-definite (there are other ways how to show the convexity but you should provide this specific one).
- (**1 Points**) We want to discriminate two Gaussians with equal covariance matrices. The log ratios of conditional probabilities are affine. Propose a feature mapping $\phi(x)$ such that for any vector w we can find a vector w' and a scalar b such that

  $$\langle w, \phi(x) \rangle = \langle w', x \rangle + b$$

  and vice versa.
- (**4 Points**) Now we discriminate Gaussians with different covariance matrices. The log ratios of conditional probabilities are quadratic. Propose a feature mapping $\phi(x)$ such that for any vector w we can find a matrix A, a vector b and a scalar c such that

  $$\langle w, \phi(x) \rangle = \langle x, Ax \rangle + \langle b, x \rangle + c$$

  and vice versa.
- (**2 Points**) Take the feature mapping from the previous point and solve Task 2. Generate 500 points from each of two Gaussians $\mathcal{N}(\mu_i, \Sigma_i)$ where

  $$\mu_1 = (0, 0)$$
  $$\Sigma_1 = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$$
  $$\mu_2 = (6, 0)$$
  $$\Sigma_2 = \begin{pmatrix} 2 & -1 \\ -1 & 7 \end{pmatrix}.$$

  You may use function `sklearn.linear_model.LogisticRegression` to solve Task (2) and `numpy.random.multivariate_normal` to generate data. The outcome of this task is a figure where the ranges $x, y \in [-20, 20]$ with the training points and the background color is the predicted class in the two-dimensional input space. Repeat the experiment multiple times, what differences you see in different runs? Explain.

Exercise 2 Linear Classification and Adversarial Training

You can solve the exercises for $2-$norm instead of $p-$norms and get full points. The solutions with $p-$norms are usually essentially the same.

In the previous assignments (such as the perceptron learning one) we wanted to find any linear classifier correctly classifying the training data. However, there are many such classifiers and we should prefer the ones that will generalize well to unseen test data. For instance, if all points are classified correctly, then we can hope that a linear classifier for which the distance of the training points to the decision boundary is large will generalize well.

At first, we consider a classification with a hinge loss

$$\max\{0, 1 - y(\langle x, w\rangle + b)\},$$

where in order to achieve zero loss, it is not sufficient to correctly classify the training examples (that is, $y(\langle x, w\rangle + b) > 0$), we also require a functional margin in order to achieve zero loss (that is, $y(\langle x, w\rangle + b) \geq 1$). In the first exercise you will show that this is not enough to enforce that the distance of the points to the decision boundary is large.

Then we look at the distance to the decision boundary induced by p-norms and show that adversarial training of the hinge loss leads to decision boundaries that are not close to the training data. (**Hint:** Hölder's inequality says that $|\langle w, x\rangle| \leq \|w\|_p \|x\|_q$ where $p, q \geq 1$ and $1/p + 1/q = 1$. For any vector w there exists a non-zero vector x such that the inequality is tight.)

a. **(3 Points)** Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \{-1, 1\}$ be the training pairs. Assume that the points are linearly separable; that is, there exists a vector w^* and a scalar b^* such that $y_i(\langle w^*, x_i\rangle + b^*) > 0$ for all i.

- Show that then there exists a pair (w, b) such that the hinge loss,

$$\frac{1}{n}\sum_{i=1}^{n} \max\{0, 1 - y_i(\langle x_i, w\rangle + b)\}, \tag{3}$$

is zero.

- Argue that the hyperplane $H = \{x \mid \langle w, x\rangle + b = 0\}$ found in the previous part can be arbitrarily close to the training data (**Hint:** we have derived the Euclidean distance of a point to a hyperplane in the lecture).

b. **(3 Points)** Let a linear classifier be given by $w \in \mathbb{R}^d \setminus \{\mathbf{0}\}$ and $b \in \mathbb{R}$. The decision boundary is the hyperplane $\{x \mid \langle w, x\rangle + b = 0\}$. Find the ℓ_p-distance of a point $a \in \mathbb{R}^d$ to the decision boundary ($p \geq 1$). That is, find the value of the optimization problem:

$$\begin{aligned} &\min_x \|a - x\|_p \\ &\text{s.t. } \langle w, x\rangle + b = 0 \end{aligned}$$

(**Hint:** Hölder inequality might be useful.)

c. **(2 Points)** Let $\varepsilon > 0$. The inner maximization problem of adversarial training is given by

$$\begin{aligned} &\max_{\delta_1,\ldots,\delta_n} \sum_{i=1}^{n} \max\{0, 1 - y_i(\langle w, x + \delta_i\rangle + b)\} \\ &\text{s.t. } \|\delta_i\|_p \leq \varepsilon, \quad \forall i. \end{aligned}$$

Provide a closed form expression for the optimal value of the objective of this problem (**Hint:** again Hölder inequality might be useful.)

d. **(2 Points)** Let $\varepsilon > 0$. Show that for the resulting linear classifier that we get by solving the adversarial training problem

$$\begin{aligned} &\min_{w,b} \max_{\delta_1,\ldots,\delta_n} \sum_{i=1}^{n} \max\{0, 1 - y_i(\langle w, x + \delta_i\rangle + b)\} \\ &\text{s.t. } \|\delta_i\|_p \leq \varepsilon, \quad \forall i. \end{aligned}$$

it holds that it separates the points and that the minimal ℓ_p-distance to the decision boundary of the training points ($p \geq 1$) is larger than ε if there exists any such linear classifier.