Aleksei Maliutin
1224****
a*****n@student.uva.nl

Homework Assignment 5
Machine Learning 1, 18/19

2018-12-16

1 A K-Sided Die

In the lecture we have discussed Lagrange multipliers and you have the opportunity to practise with Lagrange multipliers in the practice homework. Here, we will use Lagrange multipliers in a Machine Learning setting, i.e., we will derive the MAP solution for the parameters $\boldsymbol{\theta}$ of a Dirichlet-multinomial model, which is a model of the outcomes of a K-sided die. Although the this may seem like another toy example, the methods we use are often used to analyse real-world data.
Assume we have a K-sided die and we observed N dice rolls $\{x_1, ..., x_N\}$, where x_i denotes the result from the ith roll, i.e., $x_i \in \{1, ..., K\}$ for all i. Assuming iid data, we can write the likelihood as follows:

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_{x_n} = \prod_{k=1}^{K} \theta_k^{N_k} \text{ , where } N_k = \sum_{n=1}^{N} [x_n = k]$$

From multiplying the likelihood by the prior and some straightforward manipulations, we can find the posterior distribution:

$$\begin{aligned} p(\boldsymbol{\theta}|\mathcal{D}) &= \text{Dir}(\boldsymbol{\theta}|\alpha_1 + N_1, ..., \alpha_K + N_K) \\ &= \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1} \end{aligned}$$

1. (1 point)

Derive the log-posterior
ANSWER:

$$\begin{aligned} \ln p(\boldsymbol{\theta}|\mathcal{D}) = \ln \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1} = \sum_{k=1}^{K} \ln \theta_k^{N_k + \alpha_k - 1} = \\ = \sum_{k=1}^{K} (N_k + \alpha_k - 1) \cdot \ln \theta_k \end{aligned}$$

2. (1 point)

Define the Lagrangian $l(\boldsymbol{\theta}, \lambda, \boldsymbol{\mu})$, where λ is the Lagrange multiplier that corresponds to the sum-to-one constraint, and k the Lagrange multiplier that corresponds to the $\theta_k \geq 0$ constraint. Note that, although it is not strictly necessary to include the second constraint for this problem, you are required to include both for this assignment
ANSWER:
According to the task, we need to maximize the log-posterior function $\ln p(\boldsymbol{\theta}|\mathcal{D})$ with constraints $\sum_{k=1}^{K} \theta_k = 1$ and $\theta_k \geq 0$. Hence, the Lagrangian is the following:

$$l(\boldsymbol{\theta}, \lambda, \boldsymbol{\mu}) = \sum_{k=1}^{K} (N_k + \alpha_k - 1) \cdot \ln \theta_k + \lambda \cdot (\sum_{k=1}^{K} \theta_k - 1) + \sum_{k=1}^{K} \mu_k \theta_k$$

3. (2 points)

State the KKT conditions:
ANSWER:

$$\text{3K constraints } \begin{cases} \mu_k \geq 0 \\ \theta_k \geq 0 \\ \mu_k \cdot \theta_k = 0 \end{cases} , \forall k \in \{1, ..., K\}$$

4. (2 points)

Find $\boldsymbol{\theta}_{MAP}$
ANSWER:
Let's take the derivative of the Lagrangian $l(\boldsymbol{\theta}, \lambda, \boldsymbol{\mu})$ w.r.t. primal variable $\boldsymbol{\theta}$

$$\frac{\partial l(\boldsymbol{\theta}, \lambda, \boldsymbol{\mu})}{\partial \theta_k} = \frac{\partial}{\partial \theta_k} \sum_{a=1}^{K} \underbrace{(N_a + \alpha_a - 1) \cdot \ln \theta_a}_{\text{depends on } \theta_k \text{ only if } a=k} + \lambda \cdot (\sum_{a=1}^{K} \theta_a - 1) + \sum_{a=1}^{K} \mu_a \theta_a =$$
$$= \frac{N_k + \alpha_k - 1}{\theta_k} + \lambda + \mu_k = 0$$

and KKT

$$\begin{aligned} \mu_k &\geq 0 \\ \theta_k &\geq 0 \; \forall k \in \{1, ..., K\} \\ \mu_k \cdot \theta_k &= 0 \end{aligned}$$

Hence, it's obvious, that $\theta_k > 0 \; \forall k \in \{1, ..., K\}$, what leads to $\mu_k = 0 \; \forall k \in \{1, ..., K\}$, as a result, we are getting the following $\boldsymbol{\theta}_{MAP}$

$$\underbrace{\frac{N_k + \alpha_k - 1}{\theta_k}}_{\neq 0} + \lambda + \underbrace{\mu_k}_{=0} = 0$$
$$N_k + \alpha_k - 1 = -\lambda \cdot \theta_k$$
$$\theta_k = -\frac{N_k + \alpha_k - 1}{\lambda}$$
$$\boldsymbol{\theta}_{MAP} = \frac{\mathbf{N} + \boldsymbol{\alpha} - \mathbf{1}}{-\lambda}$$

2 Maximum Margin Classifier

Assume a dataset $\mathcal{D} = \{(\mathbf{x}_1, t_1), ..., (\mathbf{x}_N, t_N)\}$ where $\mathbf{x}_n \in \mathbb{R}^2$ and $t_n \in \{-1, +1\}$. Upon inspection of the data, we make the assumption that there exists a circle with radius R that separates the data (up to some exceptions). The datapoints within the circle are assigned label $t_n = -1$ and the datapoints outside of the circle are assigned label $t_n = +1$. See Figure 1 for an illustration of this assumption. Now, we do not want to find any circle that separates the data, we want to find the circle with radius R that has the maximum margin. For this assignment, we will make the simplifying assumption that the data (and thus the circle that separates the data) lies centered around the origin (0, 0).

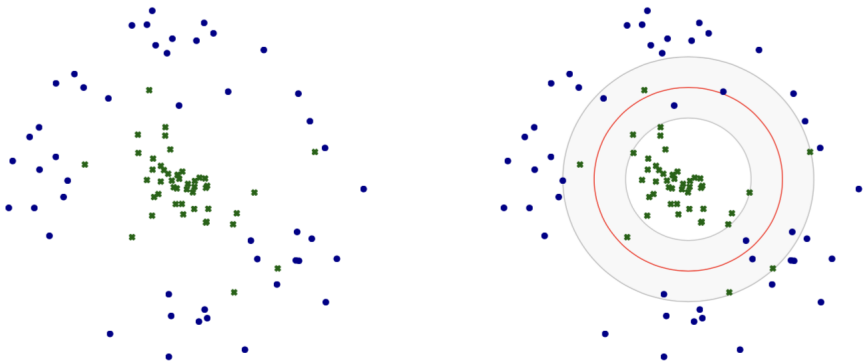

Figure 1: Illustration of data used in Assignment 3.

However, based on Figure 1, it does not seem that the data is perfectly separable, hence, we will introduce

slack variables. We will now state the primal program that will find such a circle:

$$\underset{R,\alpha,\xi}{\operatorname{argmin}} \frac{1}{2}\alpha^2 + C\sum_{n=1}^{N}\xi_n \qquad \text{s.t.} \forall n : t_n(\alpha\|\mathbf{x}_n\| - R) \geq 1 - \xi_n, \xi_n \geq 0$$

Answer the following questions:

1. (1 point)

Introduce Lagrange multipliers for the constraints and write down the primal Lagrangian. Use the following notation: $\{\lambda_n\}$ are the Lagrange multipliers for the first constraint and $\{\mu_n\}$ for the second
ANSWER:

$$L(R, \alpha, \boldsymbol{\xi}) = \frac{1}{2}\alpha^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N}\lambda_n\Big(t_n(\alpha\|\mathbf{x}_n\| - R) - 1 + \xi_n\Big) - \sum_{n=1}^{N}\mu_n\xi_n$$

2. (2 points)

How many KKT conditions are there? Write down all KKT conditions.
ANSWER:

$$\underbrace{\begin{gathered}\lambda_n \geq 0 \\ t_n(\alpha\|\mathbf{x}_n\| - R) - 1 + \xi_n \geq 0 \\ \lambda_n\Big(t_n(\alpha\|\mathbf{x}_n\| - R) - 1 + \xi_n\Big) = 0\end{gathered}}_{3N \text{ conditions}} \qquad \underbrace{\begin{gathered}\mu_n \geq 0 \\ \xi_n \geq 0 \\ \mu_n\xi_n = 0\end{gathered}}_{3N \text{ conditions}}$$

Hence, total number of KKT conditions is **6N**.

3. (3 points)

Derive the dual Lagrangian and specify the dual optimization problem.
ANSWER:
Find the derivatives of lagrangian w.r.t. primal variables:

$$\begin{aligned}
\frac{\partial}{\partial\alpha}L(R, \alpha, \boldsymbol{\xi}) = \alpha - \sum_{n=1}^{N}\lambda_n t_n\|\mathbf{x}_n\| = 0 \rightarrow \qquad & \alpha = \sum_{n=1}^{N}\lambda_n t_n\|\mathbf{x}_n\| \\
\frac{\partial}{\partial R}L(R, \alpha, \boldsymbol{\xi}) = -\sum_{n=1}^{N}\lambda_n t_n = 0 \rightarrow \qquad & 0 = \sum_{n=1}^{N}\lambda_n t_n \\
\frac{\partial}{\partial\xi_n}L(R, \alpha, \boldsymbol{\xi}) = C - \lambda_n - \mu_n = 0 \rightarrow \qquad & \lambda_n = C - \mu_n
\end{aligned}$$

Now, we can eliminate primal variables in Lagrangian:

$$\tilde{L}(\boldsymbol{\lambda}) = \frac{1}{2}\Big(\underbrace{\sum_{n=1}^{N}\lambda_n t_n\|\mathbf{x}_n\|}_{=\alpha}\Big)^2 + \sum_{n=1}^{N}(\underbrace{C - \mu_n}_{=\lambda_n})\xi_n - \sum_{n=1}^{N}\lambda_n t_n\underbrace{\sum_{n=1}^{N}\lambda_n t_n\|\mathbf{x}_n\|}_{=\alpha}\|\mathbf{x}_n\| - \underbrace{\sum_{n=1}^{N}\lambda_n t_n}_{=0}R + \sum_{n=1}^{N}\lambda_n - \sum_{n=1}^{N}\lambda_n\xi_n =$$
$$= \sum_{n=1}^{N}\lambda_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\lambda_n t_n\lambda_m t_m\|\mathbf{x}_n\|\|\mathbf{x}_m\|$$

and constraints:

$$0 \leq \lambda_n \leq C, \forall n$$
$$\sum_{n=1}^{N}\lambda_n t_n = 0, \forall n$$

4. (1 point)

Note that, because we have used a nonlinear (circular) decision boundary, we have already written down a kernelized dual Lagrangian. What is the explicit form of kernel in your final solution to the dual lagrangian?
ANSWER:

$$k(\mathbf{x}_n, \mathbf{x}_m) = \|\mathbf{x}_n\|\|\mathbf{x}_m\| = \sqrt{\mathbf{x}_n^\mathsf{T}\mathbf{x}_n \cdot \mathbf{x}_m^\mathsf{T}\mathbf{x}_m} = \sqrt{(x_{n1}^2 + x_{n2}^2) \cdot (x_{m1}^2 + x_{m2}^2)} \in \mathbb{R}^1$$

5. (1 point)

The dual program will return optimal values for $\{\lambda_n\}$. What is the minimum number of λ_n for which $0 < \lambda_n < C$ will hold?
ANSWER:
From inequality $0 < \lambda_n < C$ we can make a conclusion, that *n*-th datapoint should the closest one to the decision boundary, because based on KKT should be true the following equations:

$$\lambda_n\Big(t_n(\alpha\|\mathbf{x}_n\| - R) - 1 + \xi_n\Big) = 0 \qquad\qquad \mu_n\xi_n = 0$$

what is possible only if

$$t_n(\alpha\|\mathbf{x}_n\| - R) - 1 + \xi_n = 0 \qquad\qquad \xi_n = 0$$

Hence, if we assume that there is no such $0 < \lambda_n < C$, it means that we don't have a closest datapoint to border... what is albosulte nonsense, because at least one datapoint have to be the closest one. Therefore, the mininum number λ_n for which $0 < \lambda_n < C$ is at least 1.

6. (1 point)

Assume we have solved the dual program. Now we want to apply our maximum margin classifier to a new test case $\mathbf{x}^\star$. Describe how to classify the new datapoint $\mathbf{x}^\star$ in dual space.
ANSWER:
Prediction of class for the new datapoint $\mathbf{x}^\star$ is the sign of the following function:

$$\begin{aligned} y(\mathbf{x}^\star) = \alpha\|\mathbf{x}^\star\| - R = \sum_{n=1}^{N} \lambda_n t_n \|\mathbf{x}_n\|\|\mathbf{x}^\star\| - R = \\ = \sum_{n=1}^{N} \lambda_n t_n k(\mathbf{x}_n, \mathbf{x}^\star) - R \end{aligned}$$

7. (2 points)

Use the KKT conditions to derive wich data cases $\mathbf{x}_n$ will have $\lambda_n > 0$ and which ones will have $\mu_n > 0$.
ANSWER:
First of all, a reminder of some of KKT conditions:

$$\lambda_n\Big(t_n(\alpha\|\mathbf{x}_n\| - R) - 1 + \xi_n\Big) = 0 \qquad\qquad \mu_n\xi_n = 0$$

hence, if we will have $\lambda_n > 0$

$$\begin{aligned} \lambda_n > 0 \quad &\Rightarrow \quad t_n(\alpha\|\mathbf{x}_n\| - R) = 1 - \xi_n \\ \text{if also } \mu_n > 0 \quad &\Rightarrow \quad \xi_n = 0 \quad \Rightarrow \quad \mathbf{x}_n \text{ correctly classified, on margin (=support vector)} \\ \text{else if } \mu_n = 0 \quad &\Rightarrow \quad \xi_n \geq 0 \quad \Rightarrow \quad \begin{cases} \xi_n \leq 1 & \mathbf{x}_n \text{ correctly classified, but within margin,} \\ \xi_n > 1 & \mathbf{x}_n \text{ misclassifed datapoint} \end{cases} \end{aligned}$$

If we will have $\mu_n > 0$, it means that slack variable should be zero, in other words, that datapoint is correctly classified:

$$\begin{aligned} \mu_n > 0 \quad &\Rightarrow \quad \xi_n = 0 \quad \Rightarrow \quad \mathbf{x}_n \text{ correctly classified} \\ \text{if also } \lambda_n > 0 \quad &\Rightarrow \quad t_n(\alpha\|\mathbf{x}_n\| - R) = 1 \quad \Rightarrow \quad \mathbf{x}_n \text{ correctly classified, on margin (=support vector)} \end{aligned}$$

8. (2 points)

Compute the optimal values for the other dual variables $\{\mu_n\}$. Then, solve for the primal variables $\{\alpha, R, \boldsymbol{\xi}\}$
ANSWER:
Assume we have optimal values $\{\lambda_n^\star\}$, and using KKT we can get the following:

$$\lambda_n^\star \quad \Rightarrow \quad \lambda_n = C - \mu_n \quad \Rightarrow \quad \mu_n^\star = C - \lambda_n^\star \quad \text{optimal values for 2nd dual variable} \tag{1}$$

$$\lambda_n^\star \quad \Rightarrow \quad \alpha^\star = \sum_{n=1}^{N} \lambda_n^\star t_n \|\mathbf{x}_n\| \quad \text{optimal value for } \alpha \tag{2}$$

$$0 < \lambda_n^\star < C \quad \Rightarrow \quad t_n(\alpha \|\mathbf{x}_n\| - R) = 1 \quad \Rightarrow \quad R^\star = \frac{t_n \alpha^\star \|\mathbf{x}_n\| - 1}{t_n} \tag{3}$$

$$\lambda_n^\star = 0 \quad \Rightarrow \quad \mu_n^\star = C \quad \Rightarrow \quad \xi_n^\star = 0 \tag{4}$$

$$\lambda_n^\star = C \quad \Rightarrow \quad \mu_n^\star = 0 \quad \Rightarrow \quad \xi_n^\star = 1 - t_n(\alpha^\star \|\mathbf{x}_n\| - R^\star) \tag{5}$$

9. (1 point)

If we would use a Radial Basis Function (RBF) kernel instead of $k(\circ, \circ)$ as defined in by you in question 4, can the decision boundary be different from a circle in x-space? If yes, describe geometrically what kind of solutions we may expect when using an RBF kernel.
ANSWER:
Yes, because RBF will give us a more complex border rather than pure circle near the missclassified datapoints. Plus, depends on dataset, new border could be larger (if greed datapoints quite sparse) or smaller (if green points with high density and only a few outliers).