Machine Learning 1 - Homework 5

Available: Monday, October 7th, 2018
Deadline: Friday, 17:00 October 18th, 2018

General instructions
Unless stated otherwise, write down a derivation of your solutions. Solutions presented without a derivation that shows how the solution was obtained will not be awarded with points.

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 $\{x_1, \ldots, x_N\}$, where x_i denotes the result from the ith roll, i.e., $x_i \in \{1, \ldots, K\}$ for all i. Assuming iid data, we can write the likelihood as follows:

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

where $[c] = 1$ if the condition c holds and $[c] = 0$ otherwise (this is also known as the Iverson bracket).

Note that, since we are using a multinomial distribution as likelihood, the parameter $\boldsymbol{\theta}$ has the following constraints:

$$\sum_{k=1}^{K} \theta_k = 1, \quad \forall k : \ \theta_k \geq 0 \tag{2}$$

If we wish to introduce a prior on $\boldsymbol{\theta}$, the Dirichlet distribution is a natural choice as it is a distribution over vectors , even more so, it is also the conjugate prior to our likelihood. Hence, the prior is given by:

$$\mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}, \quad \text{for } \boldsymbol{\theta} \in S_K, \tag{3}$$

where S_K is the set of K-dimensional vectors that satisfy the constraints in (2), and $1/B(\boldsymbol{\alpha})$ is a normalization factor. From multiplying the likelihood by the prior and some

straightforward manipulations, we can find the posterior distribution:

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{4}$$

$$\propto \prod_{k=1}^{K} \theta_k^{N_k} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} = \prod_{k=1} \theta_k^{N_k + \alpha_k - 1} \tag{5}$$

From this we can identify a Dirichlet distribution, hence

$$p(\boldsymbol{\theta}|\mathcal{D}) = \mathrm{Dir}(\boldsymbol{\theta}|\alpha_1 + N_1, \ldots, \alpha_K + N_K). \tag{6}$$

Now, find the MAP estimate $\boldsymbol{\theta}_{\mathrm{MAP}}$ of $\boldsymbol{\theta}$. Don't forget to use the constraints on $\boldsymbol{\theta}$. Your answer should consist of the following steps:

1. derive the log-posterior (1 point)
2. define the **Lagrangian** $\ell(\boldsymbol{\theta}, \lambda, \boldsymbol{\mu})$, where λ is the Lagrange multiplier that corresponds to the sum-to-one constraint, and μ_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. (1 point)
3. State the KKT conditions (2 points)
4. find $\boldsymbol{\theta}_{\mathrm{MAP}}$ (2 points)

2 Maximum Margin Classifier

Assume a dataset $\mathcal{D} = \{(x_1, t_1), \ldots, (x_N, t_N)\}$ where $x_n \in \mathbb{R}^2$ and $t_n \in \{-1, +1\}$. Upon inspection of the data, we make the assumption that there exists a circle with radius $\mathcal{R}$ 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 $\mathcal{R}$ 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 $(0, 0)$.

Under the assumption that the circle perfectly separates the two classes, the distance between the decision boundary and any datapoint x_n is given by

$$t_n(||x_n|| - \mathcal{R}) = \frac{t_n(\beta||x_n|| - R)}{\beta}, \tag{7}$$

where $R = \beta\mathcal{R}$ and $\beta > 0$. Note that, the distance to the decision boundary is invariant to the rescaling $\beta \to \kappa\beta$. We can use this to set

$$t_n(\beta||x_n|| - R) = 1 \tag{8}$$

for the point x_n that is closest to the decision boundary. As such, the constraint

$$t_n(\beta||x_n|| - R) \geq 1, \qquad n = 1, \ldots, N \tag{9}$$

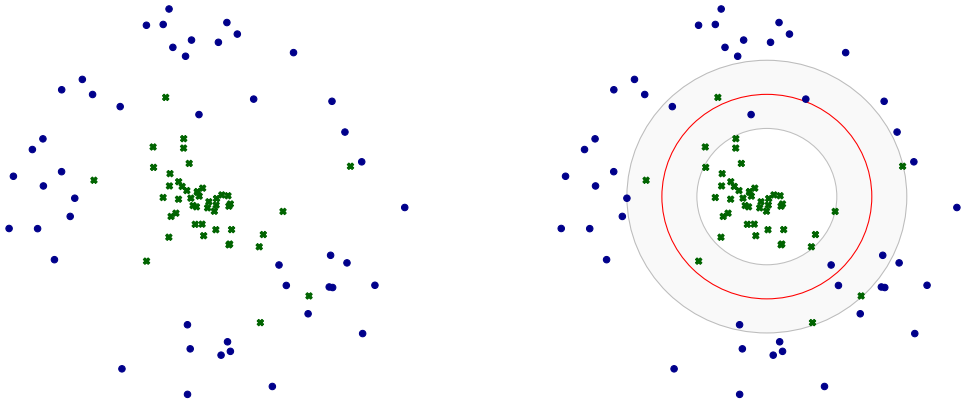

Figure 1: Illustration of data used in Assignment 3.

holds for all datapoints. 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:

$$\arg\min_{R,\beta,\boldsymbol{\xi}} \frac{1}{2}\beta^2 + C\sum_{n=1}^{N}\xi_n \qquad s.t.\ \forall n : t_n(\beta||x_n|| - R) \geq 1 - \xi_n,\ \xi_n \geq 0. \tag{10}$$

Answer the following questions:

1. 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 (1 point).

2. How many KKT conditions are there? Write down all KKT conditions (2 points).

3. Derive the dual Lagrangian and specify the dual optimization problem. That is, eliminate the primal variables $\{\beta, R, \xi_1 \ldots, \xi_N\}$ using the $\partial\ell/\partial\rho = 0$ equations, where ℓ is the primal Lagrangian and ρ is a primal variable. Do not forget to specify the constraints on the remaining dual variables (3 points).

4. Note that, because we have used a nonlinear (circular) decision boundary, we have already written down a kernelized dual Lagrangian. This results in a dual Lagrangian that no longer depends on $x_n^T x_m$ but on $\kappa(x_n, x_m)$. Also note that $\kappa(x, x') = f(x)f(x')$ is a valid kernel. What is the explicit form of $\kappa(x_n, x_m)$ in your final solution to the dual lagrangian (1 point)?

5. 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? Explain your answer (1 point).

6. Assume we have solved the dual program. Now we want to apply our maximum margin classifier to a new test case $x^\star$. Describe how to classify the new datapoint $x^\star$ in dual space (1 point).

7. Use the KKT conditions to derive wich data cases x_n will have $\lambda_n > 0$ and which ones will have $\mu_n > 0$ (2 points).

8. Compute the optimal values for the other dual variables $\{\mu_i\}$. Then, solve for the primal variables $\{\beta, R, \boldsymbol{\xi}\}$. Note that you do not need to know the dual optimization program to solve this question. You only need the KKT conditions (2 points).

9. If we would use a Radial Basis Function (RBF) kernel instead of $\kappa(\circ, \circ)$ as defined 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 (1 point).