

1. (10) **Support Vector Machines**

We are given the following dataset: $\{x_i, y_i\}$, $i = 1..N$, where each $x_i \in \mathbb{R}^d$ and $y_i = \{-1, 1\}$. Consider the primal problem formulation of a two-class SVM:

$$\begin{aligned} \text{minimize}_{\{\mathbf{w}, \xi_i, b\}} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^T \mathbf{x}_i - b) - 1 + \xi_i \geq 0 \quad \forall i \quad (1) \\ & \xi_i \geq 0 \quad \forall i \quad (2) \end{aligned}$$

a.(2 pts) Choose $C = 0$. What is the optimal value of \mathbf{w} in this case? Describe what happens to the margin in this case.

A: $\mathbf{w} = 0$. Margins become infinitely wide.

b.(2 pts) Assume that there is a hyperplane that perfectly separates the positive from the negative examples. At the optimal solution, what are the values of $\{\xi_i\}$?

A: $\xi_i = 0$, $\forall i$

Now consider the equivalent dual problem formulation:

$$\begin{aligned} \text{maximize}_{\{\alpha_i\}} \quad & \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \\ \text{subject to} \quad & \sum_i \alpha_i y_i = 0 \quad (3) \\ & 0 \leq \alpha_i \leq C \quad \forall i \quad (4) \end{aligned}$$

c.(2 pts) Consider a data-case that is located inside the margin (i.e. between the two support vector hyper-planes) at the optimal solution. What is its value for α_i ?

A: $\alpha_i = C$.

d.(2 pts) Give the two conditions that a valid kernel must satisfy.

A: *Symmetry*: $K = K^T$ and *positive semi-definiteness*: $v^T K v \geq 0$, $\forall v$

e.(2 pts) Given a kernel $K(x_i, x_j)$, and the solution to the dual problem $\{\alpha_i\}$ (assume $b = 0$), provide the equation for classifying a new test example x_{test} .

A: $y_{\text{test}} = \text{sign} \left[\sum_{i \in SV} \alpha_i y_i K(x_i, x_{\text{test}}) \right]$

2. (10 points) **Naive Bayes Classifier**

We are given the following dataset: $\{x_i, y_i\}$, $i = 1..N$, where each x_i can take one of K discrete values $k = 1, \dots, K$ and $y_i = \{1, 2, \dots, C\}$ (i.e. it's like a document with a single word k). Consider the following naive Bayes model for classification,

$$p(x_i = k | y = c) = q_{kc} \quad (5)$$

$$p(y = c) = \pi_c \quad (6)$$

a.(2 pts) Provide an expression for the joint probability: $p(x_i = k, y_i = c)$?

A: $p(x_i = k, y_i = c) = p(x_i = k | y = c) p(y = c) = q_{kc} \pi_c$

- b.(2 pts) What are the maximum likelihood estimates for the parameters of this model $\{q_{kc}, \pi_c\}$? in terms of counts. You may describe your answer in words or use an equation.
- A: $q_{kc} = \text{Nr. items with } x_i = k \text{ in class } c \text{ divided by the total Nr. of items in class } c$. $\pi_c = \text{Nr. items in class } c \text{ divided by the total Nr. of items } (N)$.
- c.(2 pts) Express the posterior probability $p(y_i = c | x_i = k)$ in terms of $\{q_{kc}, \pi_c\}$.
- A: $p(y_i = c | x_i = k) = q_{kc} \pi_c / \sum_{c'} q_{kc'} \pi_{c'}$
- d.(2 pts) Is the following statement true (explain): In the naive Bayes model, the attributes are assumed (marginally) independent.
- A: *No, conditionally independent given y*
- e.(2 pts) Explain what semi-supervised learning is. You can for instance explain what information is provided in the data (and what is not).
- A: *The labels y_i are provided for only a small subset of the data-items.*

3. (4 points) Reinforcement Learning

- a.(2 pts) Assume you are given the Q-function $Q(s, a)$ for a particular reinforcement learning problem. Express both the value function $V(s)$ and the optimal policy $\pi(s)$ in terms of the Q-function.
- A: $V(s) = \max_a Q(s, a)$, $\pi(s) = \operatorname{argmax}_a Q(s, a)$
- b.(2 pts) Consider an agent that uses Q-learning to learn a policy. Because it has no model of the environment it must experience the world and measure the state of the environment and the reward it receives for its action. While exploring the world it updates its Q-function according to $Q(s, a) \leftarrow r + \gamma \max_{a'} Q(s', a')$ where $\gamma \in (0, 1]$ the discount factor, r is the reward it receives and s' the new state of the environment. But to experience the world it needs to choose actions. The agent decides to use the policy it has learned so far (which is a function of its current Q-function – see item (a)). Explain why this strategy is NOT likely to converge to a very good policy.
- A: *The agent will not explore the world enough. It will get stuck in the first thing it tries and reinforce that policy.*

4. (6 points) Clustering

We are given data $\{x_i\}$, $i = 1, \dots, N$ with $x_i \in \mathbb{R}^d$. Consider minimizing the following K-means cost function,

$$C = \frac{1}{2} \sum_c \sum_{i \in S_c} \|x_i - \mu_c\|_{L_2}^2 \quad (7)$$

where S_c is the subset of data-items assigned to cluster c and L_2 means the L_2 norm (the same as used in class).

- a.(2 pts) Derive the K-means update rule for μ_c by computing the gradient $\partial C / \partial \mu_c$ and equating it to 0. It is implicitly assumed that the assignments are held fixed.
- A: $\partial C / \partial \mu_c = - \sum_{i \in S_c} (x_i - \mu_c) = 0 \rightarrow \mu_c = \sum_{i \in S_c} x_i / N_c$.
- b.(2 pts) Is the K-means algorithm (which alternates the above updates for $\{\mu_c\}$ with reassigning data-items to clusters) guaranteed to converge eventually? Why?
- A: *Yes, both steps are guaranteed to decrease the cost C or leave it the same, and $C \geq 0$.*

- b.(2 pts) Explain how you could use the result of the K-means algorithm to compress the dataset $\{x_i\}$, $i = 1, \dots, N$ without loss of information. You may assume that N is very large and K chosen optimally for the purpose of compression. In your answer you will have to use the fact that small real numbers can be encoded more cheaply than large real numbers. You will get half the points for explaining how you can compress the data with loss of information.
- A: *Instead of the data, you encode the cluster means and for every data-item you encode to which cluster it belongs (its assignment). This is a lossy encoding. If you also encode the hopefully small vector differences between the cluster mean and the data-item, you have not lost information, since the data-items can be perfectly reconstructed from the code. Compression is possible because the small differences can be encoded more cheaply than the original real values for $\{x_i\}$.*