
Describe the curse of dimensionality. Why does it make learning hard in high dimensional spaces?

Consider datapoints uniformly distributed in a p-dimensional unit hypercube. Define a hypercubical neighborhood around a target point with length $l$. How big must $l$ be to ensure that the hypercube contains 1% of the data (one average)? How big to contain 10%?

Consider $N$ data points uniformly distributed in a $p$-dimensional unit ball centered on the origin. Show that the median distance from the origin to the closest datapoint is given by $d(p,N) = \{1 - (1/2)^{1/N}\}^{1/p}$. Hence deduce that most datapoints are closer to the boundary of the sample space than to any other datapoint.

What do you conclude about intuitions for data in high dimensional spaces?

Hint: the probability that at least one datapoint is closer than $r$ equals one minus the probability that all datapoints are further than $r$.

Question 2. Decision Trees.

Describe how to construct a binary decision tree. What is the impurity, and decrease in impurity?

Consider a binary classification problem. The data lies in the two-dimensional plane $(x_1, x_2) \in \mathbb{R}^2$. The only positive example ($\omega = +1$) is at the origin $(0,0)$. There are 8 negative examples ($\omega = -1$) equally spaced on a circle centered on the origin with radius $R$.

Design a decision tree algorithm to classify this data. The only tests you can use are of form $f(x_1, x_2) = \text{sign}(a_1x_1 + a_2x_2 + b)$, where $a_1, a_2, b$ are constants. (Different constants give different tests). What is the smallest tree you can design to classify the data correctly? Calculate the decrease in impurity for each test in your tree.

Question 3. Primal Dual Quadratic Optimization

The primal problem is formulated as follows:

$$L_p(\bar{a}, b, \{z_i\}; \{\alpha_i, \mu_i\}) = (1/2)|\bar{a}|^2 + \gamma \sum_{i=1}^{m} z_i - \sum_{i=1}^{m} \alpha_i \{\omega_i(\bar{a} \cdot \vec{x}_i + b) - (1 - z_i)\} - \sum_{i=1}^{m} \mu_i z_i.  \quad (1)$$

Explain the meaning of all the terms and variables in this equation. What constraints do the variables satisfy? Calculate the form of the solution $\bar{a}$ by minimizing $L_p$ with respect to $\bar{a}, b, \{z_i\}$. What are the support vectors?

Obtain the dual formulation by eliminating $\bar{a}, b, \{z_i\}$ from $L_p$. Describe a strategy for solving the primal problem.
**Question 4. VC Dimension.**

Define the VC dimension $h$ of a set of classifiers. What does it mean “to shatter” the data? Obtain the VC dimension for a set of hyperplane classifiers for data lying in two dimensions. Illustrate how many points can be shattered by hyperplane classifiers (in two dimensions).

What does “probably approximately correct” mean? How does VC theory bound the risk $R[\alpha]$ in terms of the empirical risk and a capacity term? What factors does the capacity term depend on? Plot the capacity term as a function of the number of data samples $n$ for different values of these factors (3 plots are sufficient).

**Question 5. SVM.**

Consider a binary classification problem where the data is one-dimensional. There are two negative ($\omega = -1$) examples at $x = \pm 1$ and one positive ($\omega = 1$) example at $x = 0$. Show that you cannot classify this data perfectly by linear separation (i.e. by a decision rule $\text{sign}(ax + b)$ for some $a, b$).

Now formulate this problem with slack variables $\{z_i\}$. The classifier with the largest margin is obtained by solving the primal problem:

$$L_P = \frac{1}{2}a^2 + \gamma \sum_{i=1}^{3} z_i - \sum_{i=1}^{3} \alpha_i \left\{ \omega_i (ax_i + b) - (1 - z_i) \right\} - \sum_{i=1}^{3} \mu_i z_i,$$

where $\gamma$ is a constant, the $\{\alpha_i, \mu_i\}$ are Lagrange multipliers (constrained to be non-negative) and the $\{z_i\}$ are non-negative.

Minimize $L_p$ by searching for the minimum of $(1/2)a^2 + \gamma(z_1 + z_2 + z_3)$ subject to the constraints $\omega_i(ax_i + b) - (1 - z_i) \geq 0, \forall i \in \{1, 2, 3\}$ with $z_i \geq 0, \forall i \in \{1, 2, 3\}$. (Hint: exploit structure of the problem to guess where the decision boundary should be). What are the support vectors?

**Question 6.**

Show that you can solve the classification problem of Question 5 using kernel methods. (Hint: consider a special choice of kernel).

Consider a binary classification problem. The data lies in the two-dimensional plane $(x_1, x_2) \in \mathbb{R}^2$. The only positive example ($\omega = +1$) is at the origin $(0, 0)$. The are 8 negative examples ($\omega = -1$) equally spaced on a circle centered on the origin with radius $R$.

Choose a three-dimensional feature vector which enables the positive and negative examples to be separated by a linear hyperplane (in feature space).

Use these feature vectors to calculate a kernel $K(x_1, x_2; x'_1, x'_2)$ for this problem.

Show that the classifier in feature space can be expresses using kernels as $\text{sign}(g(x_1, x_2))$ where $g(x_1, x_2) = \sum_{\mu=1}^{8} \alpha_\mu \omega_\mu K(x_1, x_2; x'_1, x'_2) + \alpha_0 \omega_0 K(x_1, x_2; 0, 0) + b$, where $\{(x'_1, x'_2) : \mu = 1, ..., 8\}$ are the data points on the circle.

What is the minimal number of $\alpha$’s that need to be non-zero to enable classification? (The answer will depend on your choice of feature vectors).