\section{\(\nu\)-SVM}

The \(C\) parameter in support vector machines is not particularly intuitive. An alternative formulation is the \(\nu\)-SVM, where a parameter \(\nu \in [0,1]\) is used instead of \(C \in \mathbb{R}^+\):

\[
\min_{w, \rho} \frac{1}{2} \|w\|^2 - \nu \rho + \frac{1}{n} \sum_{i=1}^{n} \xi^i
\]

subject to

\[y^i \langle w, x^i \rangle \geq \rho - \xi^i, \text{ for } i = 1, \ldots, n\]
\[\xi^i \geq 0, \text{ for } i = 1, \ldots, n\]
\[\rho \geq 0.\]

(Note: \(\nu\) is fixed in advance, we do not optimize over it)

a) Assume a solution with \(\xi^i = 0\) for \(i = 1, \ldots, n\) exists. What’s the margin of separation between the classes?

For a general solution, let \(R^\rho\) be the fraction of margin errors on the training set:

\[
R^\rho(w) := \frac{1}{n} \left| \left\{ i : y^i \langle w, x^i \rangle < \rho \right\} \right|.
\]

b) Show: \(\nu\) bounds the fraction of margin errors we make, i.e. at the optimal solution we have \(R^\rho^*(w^*) \leq \nu\).

\section{Combining Kernels}

There are many more ways to create new kernels than we saw in the lecture:

a) Prove: for any non-negative constants, \(a^1, \ldots, a^m \geq 0\), and any positive definite kernels, \(k^1, \ldots, k^m\), the weighted linear combination \(\sum_i a^i k^i\) is a positive definite kernel.

b) Prove: for any positive definite kernel \(k(x, x')\), the function \(e^{k(x, x')}\) is also a positive definite kernel.

\section{K-means}

For data points \(x^1, \ldots, x^n\), the objective of \(K\)-means is

\[
\sum_{i=1}^{n} \|x^i - c_{z^i}\|^2
\]

where \(z^i \in \{1, \ldots, K\}\) is the cluster assignment of the example \(x^i\).

a) Show: both steps of the \(K\)-means algorithm reduce the objective or keep it constant.

b) Show: in the globally minimum of the \(K\)-means objective, there no empty clusters exist.
4 Dimensionality Reduction

Let \( \mathcal{D} \) be a dataset of an equal amount of points \( x = \left( \frac{a}{±1} \right) \) and \( x = \left( \frac{±1}{a} \right) \) with \( a \) chosen uniformly randomly for each sample in the interval \([-1, 1]\) and the signs of the other coordinates also chosen uniformly at random, \( p(+1) = p(-1) = \frac{1}{2} \).

a) Draw \( \mathcal{D} \) (qualitatively)
b) What is mean of \( \mathcal{D} \) (qualitatively)?
c) What does the covariance matrix of \( \mathcal{D} \) (qualitatively) look like?
d) What result will PCA give you if reduce from 2 to 1 dimension?

Let \( \mathcal{D}' \) be the result of applying the polar-coordinate change of coordinates to \( \mathcal{D} \).

e) Draw \( \mathcal{D}' \) (qualitatively)
f) What is the mean of \( \mathcal{D}' \) (qualitatively)?
g) What does the covariance matrix of \( \mathcal{D}' \) (qualitatively) look like?
h) What result will PCA on \( \mathcal{D}' \) give you if reduce from 2 to 1 dimension?

5 Number of Clusters

Automatically selecting the “correct” number of clusters for a clustering method is an unsolved, and potentially unsolvable problem.

a) Why can’t we treat the number of clusters just as an unknown parameter and minimize the \( K \)-means objective over it.

One suggestion for fixing this problem is the Bayesian Information Criterion (BIC), where one adds a term \( K \log D \) to the objective (\( K \) is the number of clusters, \( D \) is the dimensionality of the data).

b) Discuss this suggestion in light of the concept of regularization.

The problem of ”how many clusters” is inherently coupled to the question of ”scale”.

c) Construct a dataset that has 2 clusters if looking at a coarse scale, but 4 cluster on a finer scale.

d) Does the BIC help us to pick one of the two possibilities?

6 Practical Experiments

Implement the \( K \)-means algorithm for clustering and the \textit{EM-algorithm} for learning Gaussian Mixture Models. Create two datasets consisting of 100 random samples from each of three Gaussian in \( \mathbb{R}^3 \) with parameters:

a) \( \mu_1 = (1, 0, 0), \Sigma_1 = \text{Id}, \mu_2 = (0, 1, 0), \Sigma_2 = \text{Id}, \mu_3 = (0, 0, 1), \Sigma_3 = \text{Id} \).

b) \( \mu_1 = (0, 0, 0), \Sigma_1 = \text{diag}(1, 1, 10), \mu_2 = (0, 0, 0), \Sigma_2 = \text{diag}(1, 10, 1), \mu_3 = (0, 0, 0), \Sigma_3 = \text{diag}(10, 1, 1) \).

Apply both methods to cluster each of the datasets into 3 clusters. What results do you expect, what results do you get?

7 Entertainment I

Search on Google for the New York Times article ”How Many Computers to Identify a Cat?”. Discuss the cat neuron that they present in light of the ”1533” exercise and the lecture’s discussion on overfitting and interpreting the training error.

8 Entertainment II

Visit the Topic browser of the science magazine, http://topics.cs.princeton.edu/Science/. Discuss how the automatically found topics differ from the topic structure a human editor would have selected.