MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering and Computer Science
6.036—Introduction to Machine Learning
Spring Semester 2017

Assignment 5, Issued: Monday, April 10, Due Date: Friday, April. 21 at 9am

VC Dimension

1. We consider here a set of non-linear classifiers over scalars $x \in \mathbb{R}$. The classifiers in this set are based on a polynomial expansion of x. In particular, let $\mathcal{P}_d$ be the set of all polynomial classifiers of degree at most d, consisting of all functions $p(x) = a_d x^d + a_{d-1}x^{d-1} + \ldots + a_1 x + a_0$ with real a_i's. A $p(x)$ of this form classifies each $x \in \mathbb{R}$ into classes $+$ or $-$ according to its sign.

   (a) Consider the polynomial classifier $p(x) = x^2 - x - 1 \in \mathcal{P}_2$. What is each point $x = 0, 1, 2, 3$ classified as (i.e. positive or negative) under this classifier?

   (b) Prove that $\mathcal{P}_d$ has the VC-dimension of exactly $d + 1$. One way to approach this is by splitting the problem into two parts, first showing that the VC-dimension $\leq d + 1$ and VC-dimension $\geq d + 1$.[1]

   (c) Consider the set of monic degree d polynomials. This is the subset of $\mathcal{P}_d$ where the leading coefficient $a_d = 1$. Is the VC-dimension any different for this set?

   (d) **(optional)** Now consider the set $\mathcal{P}_d^+$ which consists of polynomial where all coefficients $a_d, \ldots, a_0 > 0$. What is the VC-dimension of this set?

   (e) **(optional)** Let $S \subseteq \{0, 1, 2, 3, \ldots, d\}$ be an arbitrary subset of non-negative integers up to d. Let $\mathcal{P}_S$ be the set of polynomials where a_i is allowed to be non-zero only if $i \in S$. Can you determine the VC-dimension of $\mathcal{P}_S$?

Information Criteria

2. Complexity measures such as the VC dimension can be used to guide the selection of different families of classifiers, e.g., whether to use one type of kernel vs another. However, we will also need a criterion to decide between one type of probability model over another, e.g., numbers of mixture components in a Gaussian mixture model.

   The Bayesian information criterion (BIC) is one such criterion. It attempts to capture the tradeoff between the log-likelihood of the data and the number of parameters that the model uses. In other words, it tries to balance how well we can fit the data, and how well we would expect to fit the data just because we have more parameters to tune. The BIC of a model M is defined as:

$$\text{BIC}(M) = l - \frac{1}{2} p \log n$$

[1] Hint: To gain further intuition for the second part, first consider the case of small dimensions $d = 0, 1, 2, \ldots$. Once you understood these cases, try doing an induction on the dimension for showing that VC-dimension is $\geq d + 1$.

where l is the highest log-likelihood of the data under the parameters of the current model, p is the number of adjustable parameters, and n is the number of data points. These information criteria reward a larger log-likelihood, but penalize the number of parameters used to train the model. In a situation where we wish to select models, we want a model with the highest BIC.

You are consulting for Data Technologies Firm (DTF), a YC data analytics startup that applies scalable machine learning AI to real-time visualizations in the cloud. They are currently evaluating a model M_1 in comparison to alternatives using the BIC.

(a) Currently, DTF processes millions of data points n as part of their model with p parameters. The Harvard grad on the team notes that since l is typically negative and obtained as a sum over data points, they could increase BIC score by reducing n. Is this a good idea?

(b) DTF is considering switching to another, more complex, model M_2, which has 5 times more parameters as their M_1. This model seems to better fit the data they have available. How much of a better fit of the data should M_2 provide in order for you to recommend a switch? Please provide an expression for the minimal difference in data log-likelihood ($\delta l = l_2 - l_1$) that would justify this change.

(c) Another information criterion is the AIC (the Aikake information criterion), defined as AIC$(M) = l - p$. A common observation is that the BIC favors simpler models than AIC. Provide a short explanation as to why this is so.

K-means and K-medoids

3. Euclidean distance is not the only way to measure the distance between two d-dimensional vectors. There is a related family of distance measures, known as l_p norms, parameterized by $p \geq 1$.

The l_p norm of a vector is defined as: $\|x\|_p = \left(\sum_j |x_j|^p\right)^{1/p}$.

The standard Euclidean distance is the l_2 norm of vector difference between the two points,
$\|x - y\|_2 = \left(\sum_j |x_j - y_j|^2\right)^{1/2}$

The Manhattan distance is the l_1 norm defined as:
$\|x - y\|_1 = \sum_j |x_j - y_j|$

Assume we have a 2D dataset consisting of $(0, -6), (4, 4), (0, 0), (-5, 2)$, $k = 2$, and we initialize the cluster centers with $(-5, 2), (0, -6)$. For this small dataset, in choosing between two equally valid exemplars for a cluster in k-medoids, choose them with priority in the order given above (i.e. all other things being equal, you would choose $(0, -6)$ as a center over $(-5, 2)$). After the algorithm converges, what will the clusters and cluster centers be if we run:

(a) K-medoids algorithm with l_1 norm

(b) K-medoids algorithm with l_2 norm

(c) K-means algorithm with l_1 norm

EM algorithm

4. Consider the following mixture of two Gaussians: $P(x; \theta) = \pi_1 N(x; \mu_1, \sigma_1^2) + \pi_2 N(x; \mu_2, \sigma_2^2)$

which has adjustable parameters $\theta = \{\pi_1, \pi_2, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\}$ (the mixing proportions, means, and variances of each Gaussian). We initialize the mixture parameters according to $\pi_1 = \pi_2 = 0.5, \mu_1 = 6, \mu_2 = 7, \sigma_1 = 1, \sigma_2 = 2$.

We have the following samples of x: $x^{(0)} = -1, x^{(1)} = 0, x^{(2)} = 4, x^{(3)} = 5, x^{(4)} = 6$, denoted collectively as the data D.

(a) Write down the expression for the log-likelihood of the data $l(D;\theta)$ under the mixture model with parameters θ.

(b) The difficulty with maximizing $l(D;\theta)$ is that the Gaussians cannot be set independently of each other; indeed, they are chosen so that they work well together to account for the data D. We could try to maximize $l(D;\theta)$ with a gradient ascent algorithm. This would require us to tune the learning rate and to make sure that each gradient step would keep the parameters within appropriate limits (e.g., mixing proportions specify a distribution).

The EM algorithm is generically available for estimating probability models with latent variables and it does not require learning rates. The algorithm optimizes a lower bound on the log-likelihood thus iteratively pushing the data likelihood upwards. The iterative algorithm is specified by two steps applied successively

E-step: infer component assignments (complete the data)

$$p(y|i) := P(y \mid x^{(i)};\theta), \text{ for } y = 1, 2, \text{ and } i = 0, \ldots, 4.$$

M-step: maximize the expected log-likelihood

$$\tilde{l}(D;\theta) := \sum_i \sum_y p(y|i) \log \frac{P(x^{(i)}, y;\theta)}{p(y|i)}.$$

with respect to θ while keeping $p(y|i)$ fixed.

If we indeed set $p(y|i)$ as in the E-step, show that the lower bound $\tilde{l}(D;\theta)$ is exactly the log-likelihood $l(D;\theta)$ before we start changing θ in the M-step. (we will omit the part that $\tilde{l}(D;\theta)$ is indeed always a lower bound)

(c) Now that we understand the algorithm, let's go back to our specific example. In the first E-step, which examples are more likely to be (but not entirely) assigned to the second Gaussian? In other words, what are the points for which $P(y = 2|x^{(i)}, \theta_0) > P(y = 1|x^{(i)}, \theta_0)$?

(d) In the first M-step, in which direction will the two Gaussians move?

(e) In the first M-step, do the variances σ_1^2 and σ_2^2 increase or decrease?

(f) Where will the two Gaussians be after we run the EM algorithm until convergence? Which of the resulting variances is larger?