

Comp 6321 - Machine Learning - Assignment 3

Federico O'Reilly Regueiro

November 10th, 2016

Question 1: Midterm preparation question

Propose an adequate learning algorithm for each instance.

- 1.a 1000 samples, 6-dimensional continuous space, classify ~ 100 examples.
- 1.b Classifier for children in special-ed, justified to the board before it's implemented.
- 1.c Binary classification of 1 million bits (empirical preference rate for others), very large data-set. Frequent updates.
- 1.d 40 attributes, discrete and continuous, some have noise; only about 50 labeled observations.

Question 2: Properties of entropy

2.a Compute the following for (X, Y) :

$$p(0, 0) = 1/3, p(0, 1) = 1/3, p(1, 0) = 0, p(1, 1) = 1/3.$$

- i $H[x] = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) = .9182$
- ii $H[y] = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) = .9182$
- iii $H[y|x] = \sum_x p(x)H[Y|X=x] = \frac{2}{3}\left(-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) = \frac{2}{3}$
- iv $H[x|y] = \sum_y p(y)H[X|Y=y] = \frac{2}{3}\left(-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) = \frac{2}{3}$
- v $H[x, y] = 3\left(-\frac{1}{3}\log_2\left(\frac{1}{3}\right)\right) = -\log_2\left(\frac{1}{3}\right) = 1.5849$
- vi $I[x, y] = \sum_x \sum_y p(x, y)\log_2\left(\frac{p(x, y)}{p(x)p(y)}\right) = H[x] - H[x|y] = 0.2516$

2.b Prove maximum entropy in a discrete distribution happens in U

We wish to find:

$$\arg \max_{p_n} \sum_{n=1}^N p_n \log(p_n)$$

With constraints:

$$1 - \sum_{n=1}^N p_n = 0$$

We use a Lagrangian multiplier such that:

$$\nabla_{p_1, p_2, \dots, p_N} \sum_{n=1}^N p_n \log(p_n) = \nabla_{p_1, p_2, \dots, p_N} \lambda (1 - \sum_{n=1}^N p_n)$$

We are thus left with a system:

$$\begin{aligned} \frac{\partial}{\partial p_1} \sum_{n=1}^N p_n \log(p_n) &= \frac{\partial}{\partial p_1} \lambda (1 - \sum_{n=1}^N p_n) \\ \frac{\partial}{\partial p_2} \sum_{n=1}^N p_n \log(p_n) &= \frac{\partial}{\partial p_2} \lambda (1 - \sum_{n=1}^N p_n) \\ &\vdots \\ \frac{\partial}{\partial p_N} \sum_{n=1}^N p_n \log(p_n) &= \frac{\partial}{\partial p_N} \lambda (1 - \sum_{n=1}^N p_n) \\ 1 - \sum_{n=1}^N p_n &= 0 \end{aligned}$$

Which in turn yields:

$$\begin{aligned} \log(p_1) + 1 &= \lambda p_1 \\ \log(p_2) + 1 &= \lambda p_2 \\ &\vdots \\ \log(p_N) + 1 &= \lambda p_N \\ 1 - \sum_{n=1}^N p_n &= 0 \end{aligned}$$

From which it is clear that $p_1 = p_2 = \dots p_N = \frac{1}{N}$, which is precisely a discrete uniform distribution.

2.c Show that T_1 wins

The notes show two possible tests for a decision tree. T_1 , where the left child has $[20+, 10-]$ possible outcomes in its sub-trees and the right node has $[10+, 0-]$. T_2 , on the other hand, yields: $left = [15+, 7-]$; $right = [15+, 3-]$.

The best choice should yield the maximum information gain $I[p, T_n], n \in \{1, 2\}$. So for T_1 :

$$\begin{aligned} H[p] &= -\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right) = 0.8112 \\ H[p|T_1 = t] &= -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right) = 0.9182 \\ H[p|T_1 = f] &= 0 \\ H[p|T_1] &= p(T_1 = t)H[p|T_1 = t] + p(T_1 = f)H[p|T_1 = f] \\ &= 0.6887 \\ I[p, T_1] &= H[p] - H[p|T_1] = 0.1225 \end{aligned}$$

Whereas for T_2 we have:

$$\begin{aligned} H[p|T_2 = t] &= -\frac{15}{22}\log_2\left(\frac{15}{22}\right) - \frac{7}{22}\log_2\left(\frac{7}{22}\right) = 0.9024 \\ H[p|T_2 = f] &= -\frac{15}{18}\log_2\left(\frac{15}{18}\right) - \frac{3}{18}\log_2\left(\frac{3}{18}\right) = 0.65002 \\ H[p|T_2] &= p(T_2 = t)H[p|T_2 = t] + p(T_2 = f)H[p|T_2 = f] \\ &= \frac{22}{40}0.9024 + \frac{18}{40}0.65002 = 0.7888 \\ I[p, T_2] &= H[p] - H[p|T_2] = 0.02245 \end{aligned}$$

From which we can see that we gain much more information from knowing the result of T_1 than by knowing the result of T_2 .

Question 3: Kernels

Suppose $k_1(\mathbf{x}, \mathbf{z})$ and $k_2(\mathbf{x}, \mathbf{z})$ are valid kernels over $\mathbb{R}^n \times \mathbb{R}^n$. Prove or disprove that the following are valid kernels.

Use Mercer's theorem regarding the kernel or Gram matrix or the fact that a kernel can be expressed as $k(x, z) = \phi(\mathbf{x})^T \phi(\mathbf{z})$.

preliminaries

From Mercer, we know for each $k_1(\mathbf{x}, \mathbf{z})$ and $k_2(\mathbf{x}, \mathbf{z})$ we have corresponding kernel matrices \mathbf{M}_1 and \mathbf{M}_2 which are symmetric and positive semi-definite.

For both \mathbf{M}_1 and \mathbf{M}_2 :

Symmetry:

$$\mathbf{M}_i = \mathbf{M}_i^T \tag{1}$$

Positive semidefiniteness:

$$\mathbf{x}^T \mathbf{M}_i \mathbf{x} \geq 0 \quad (2)$$

$$|\mathbf{M}_i| \geq 0 \quad (3)$$

3.a $k(\mathbf{x}, \mathbf{z}) = ak_1(\mathbf{x}, \mathbf{z}) + bk_2(\mathbf{x}, \mathbf{z}), a, b > 0; a, b \in \mathbb{R}$

Firstly, we establish that for any valid kernel $k(\mathbf{x}, \mathbf{z}), ak(\mathbf{x}, \mathbf{z})|a > 0; a \in \mathbb{R}$: We know that for a square matrix \mathbf{A} of size $n \times n$, $|a\mathbf{A}| = a^n |\mathbf{A}|$, and since $a^n \geq 0 \forall n \in \mathbb{N}, a > 0$ Then the property from equation 3 holds for both of our summands. Additionally, since the scalar multiplication of a symmetric matrix yields another symmetric matrix, both summands are valid kernels.

Now, let us say:

$$ak_1(\mathbf{x}, \mathbf{z}) = k'_1(\mathbf{x}, \mathbf{z})$$

and

$$bk_2(\mathbf{x}, \mathbf{z}) = k'_2(\mathbf{x}, \mathbf{z})$$

are both valid kernels with kernel matrices \mathbf{M}'_1 and \mathbf{M}'_2 . The addition of two symmetric matrices yields a symmetric matrix, so we need to check for positive semi-definiteness.

Since both \mathbf{M}'_1 and \mathbf{M}'_2 are symmetric we can write:

$$\begin{aligned} \mathbf{M}'_1 &= \mathbf{U}^T \mathbf{U} \\ \mathbf{M}'_2 &= \mathbf{V}^T \mathbf{V} \end{aligned}$$

and using equation 2:

$$\begin{aligned} (\mathbf{x}^T \mathbf{U}^T \mathbf{U} \mathbf{x} + \mathbf{x}^T \mathbf{V}^T \mathbf{V} \mathbf{x}) &\geq 0 \\ \mathbf{x}^T (\mathbf{U}^T \mathbf{U} + \mathbf{V}^T \mathbf{V}) \mathbf{x} &\geq 0 \\ \mathbf{x}^T (\mathbf{M}'_1 + \mathbf{M}'_2) \mathbf{x} &\geq 0 \end{aligned}$$

Which proves that $k(\mathbf{x}, \mathbf{z}) = ak_1(\mathbf{x}, \mathbf{z}) + bk_2(\mathbf{x}, \mathbf{z}), a, b > 0; a, b \in \mathbb{R}$ is a valid kernel.

3.b $k(\mathbf{x}, \mathbf{z}) = ak_1(\mathbf{x}, \mathbf{z}) - bk_2(\mathbf{x}, \mathbf{z}), a, b > 0; a, b \in \mathbb{R}$

Suppose:

$$a = 1, b = 1, \mathbf{M}_1 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \mathbf{M}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

Both \mathbf{M}_1 and \mathbf{M}_2 symmetric, positive semi-definite matrices. Yet $\mathbf{M}' = a\mathbf{M}_1 - b\mathbf{M}_2$ would yield:

$$\mathbf{M}' = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

The eigenvalues of which are $\lambda_1 = -1, \lambda_2 = 1$, making \mathbf{M}' a non positive semi-definite matrix and thus $k(\mathbf{x}, \mathbf{z})$ is not a valid kernel.

3.c $k(\mathbf{x}, \mathbf{z}) = k_1(\mathbf{x}, \mathbf{z})k_2(\mathbf{x}, \mathbf{z})$

The kernel matrix \mathbf{M}' of the product of two matrices $k_1(\mathbf{x}, \mathbf{z}), k_2(\mathbf{x}, \mathbf{z})$ is equivalent to the element-wise multiplication of the respective two kernel matrices $\mathbf{M}_1, \mathbf{M}_2$. This is also known as the Hadamard product or the Schur product. The Schur product theorem states that the Schur product of two positive semi-definite matrices is also positive semi-definite. It is trivial to show that symmetry is preserved under such conditions.

3.d $k(\mathbf{x}, \mathbf{z}) = f(\mathbf{x})f(\mathbf{z}), \text{ where } f : \mathbb{R}^n \rightarrow \mathbb{R}$

Here we rely on the fact that a kernel can be expressed as $k(x, z) = \phi(\mathbf{x})^T \phi(\mathbf{z})$ where $\phi(\mathbf{x})$ maps \mathbf{x} onto an n -dimensional space.

It is trivial to see that if $n = 1$ and $\phi = f$, $f(\mathbf{x})f(\mathbf{z})$ constitutes a valid kernel.

3.e $k(\mathbf{x}, \mathbf{z}) = f(\mathbf{x})f(\mathbf{z}), \text{ where } p \text{ pdf.}$

The same rationale as question 3.d applies here.

Question 4: Nearest neighbour vs decision trees, do boundaries coincide?

Boundaries do not necessarily coincide for these two classification strategies. Boundaries for trees are composed of hyper-planes that are orthogonal to the feature(s) chosen for the separation and pass through the midpoint between neighboring points along the axis of the chosen feature(s); so their boundaries can have one direction out of a set number of directions for a given space. Conversely, boundaries for nearest-neighbours correspond to a Voronoi tessellation, where the space is populated with hyper-planes running orthogonal to lines uniting nearest neighbors (thus said hyperplanes can have an arbitrary direction) and passing through the midpoint between neighbours along this axis.

To illustrate: Suppose an

Question 5: Bayes rate

For the following univariate case where $P(\omega_i) = \frac{1}{c}$ and

$$P(x|\omega_i) = \begin{cases} 1 & 0 \leq x \leq \frac{cr}{c-1} \\ 1 & i \leq x \leq i+1 - \frac{cr}{c-1} \\ 0 & \text{otherwise} \end{cases}$$

5.a Show that $P^* = r$

The minimal multi-class classification error rate P^* is given by:

$$P^* = 1 - \int \arg \max_i P(\omega_i) P(x|\omega_i) dx$$

And given the class density and probability, we can see that for any region with overlapping densities, the choice of any i will maximize. Thus:

$$\begin{aligned} P^* &= 1 - \int P(\omega_1) P(x|\omega_1) dx \\ &= 1 - \frac{1}{c} \int_0^{\frac{cr}{c-1}} dx - \sum_{i=1}^c \frac{1}{c} \int_i^{i+1 - \frac{cr}{c-1}} dx \\ &= 1 - \left(\frac{r}{c-1} + 1 - \frac{cr}{c-1} \right) \\ &= 1 - \left(\frac{r + (c-1) - cr}{c-1} \right) \\ &= 1 - \left(\frac{r(c-1) - (c-1)}{c-1} \right) \\ &= 1 + r - 1 \\ &= r \end{aligned}$$

If, however, $r < 0, |r| < \frac{c-1}{c}$, there are $c-1$ density overlaps. Thus we would need to develop otherwise:

$$\begin{aligned} P^* &= 1 - \int P(\omega_1) P(x|\omega_1) dx \\ &= \sum_{i=1}^{c-1} \frac{1}{c} \int_{i+1}^{i+1 + \frac{c|r|}{c-1}} dx \\ &= \left(\frac{c-1}{c} \frac{|r|}{c-1} \right) \\ &= |r| \end{aligned}$$

5.b Show the nearest-neighbor rate $P = P^*$

$$\begin{aligned}
LNN &= \int \left[1 - \sum_{i=1}^c P^2(\omega_i|x) \right] p(x) dx \\
&= \int \left[1 - \sum_{i=1}^c \left(\frac{P(x|\omega_i)P(\omega_i)}{p(x)} \right)^2 \right] p(x) dx \\
&= \int p(x) - \sum_{i=1}^c \frac{P(x|\omega_i)^2 P(\omega_i)^2}{p(x)} dx \\
&= \int p(x) - \sum_{i=1}^c \frac{P(x|\omega_i)P(\omega_i)(P(x|\omega_i)P(\omega_i))}{p(x)} dx \\
&= \int p(x) - \sum_{i=1}^c \frac{P(x|\omega_i)P(\omega_i)p(x)}{p(x)} dx \\
&= \int p(x) dx - \frac{1}{c} \int_0^{\frac{cr}{c-1}} dx - \sum_{i=1}^c \frac{1}{c} \int_i^{i+1-\frac{cr}{c-1}} dx \\
&= 1 - \frac{1}{c} \frac{cr}{c-1} - 1 - \frac{cr}{c-1} \\
&= \frac{cr-r}{c-1} \\
&= r
\end{aligned}$$

Question 6: Implementation