50.039 – Theory and Practice of Deep Learning

Alex

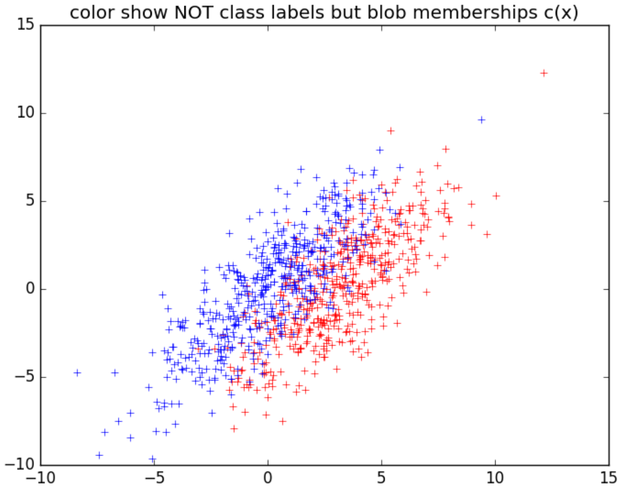
Week 01: Discriminative ML - quick intro

[The following notes are compiled from various sources such as textbooks, lecture materials, Web resources and are shared for academic purposes only, intended for use by students registered for a specific course. In the interest of brevity, every source is not cited. The compiler of these notes gratefully acknowledges all such sources.]

1 Theory homework: a known $P(x, y)$

1.1 Data generation idea

We want to generate data for classification with 2 classes. We need pairs of data x and label y. We assume two classes: $y \in \{0, 1\}$. We assume the data being 2-dim: $x \in \mathbb{R}^d, d = 2$. The coarse idea of how to generate data is for this exercise is: **we will draw data from 2 gaussian blobs. Depending on whether the data is from blob 1 or blob 2, the probability of having a label $y = 0$ will be different.**



1.2 Drawing algorithm

Repeat for n data pairs (x, y)

- draw a random value for the membership variable $C \in \{1,2\}$. $P(C = 1) = 0.5$
- draw x from a gaussian with index being equal to the value of C. If $C = 2$, then draw from gaussian with index 2.
- using the value of C, draw y according to

$$p(y = 0|x, c(x) = 1) = 0.2$$
$$p(y = 0|x, c(x) = 2) = 0.7$$

1.3 Homework and Theory part: What is the distribution of (x, y)

What is the distribution of (x, y)? It is important to understand here: x has a density, y has a discrete probability.

Our distribution of (x, y) depends on whether they come from gaussian 1 or from gaussian 2, and coming from one of the gaussians is a disjoint event, so we can write:

$$p(y, x) = p(y, x, \{c(x) = 1\} \text{ or } \{c(x) = 2\}) = ???$$

Homework task:

Goal: to understand how $p(x, y)$ looks like when data is generated from 2 (or k) clusters of (x, y) such that for every cluster x follows some distribution and the distribution of y depends only on the cluster index $c(x) \in \{1, 2\}$

Write down the expression for $p(x, y)$ as a function of:

- $P(c(x) = 1), P(c(x) = 2)$ - which is the probability to draw a data point x from a cluster
- $f(x|c(x) = 1), f(x|c(x) = 2)$ - which is the distribution of the datapoints, given that they come from a particular cluster ,
- and of $p(y = 0|x, c(x) = 1), p(y = 0|x, c(x) = 2)$.
- for the first result use above probability symbols, do not plug anything in.
- then plug in the values that you have, use for the homework

$$p(y = 0|x, c(x) = 1) = 0.2$$
$$p(y = 0|x, c(x) = 2) = 0.7$$

  and $P(C = 1) = 0.5$. Do not plug in any parameters for the density (thats just gory notation).

Note that we assume here, that the distribution of y depends only on the cluster membership $c(x)$ and not on the value of the data point x itself, that is:

$$p(y = 0|x, c(x)) = p(y = 0|c(x))$$

Note also:

$$p(y = 0|x, c(x) = 1) + p(y = 1|x, c(x) = 1) = 1$$
$$p(y = 0|x, c(x) = 2) + p(y = 1|x, c(x) = 2) = 1$$

You can start from above equation.

2 Theory homework: a matrix can be seen as just a vector. Blue pill = red pill.

An inner product between two D-dim vectors v, w is defined by

$$v \cdot w = \sum_{d=1}^{D} v_d w_d \tag{1}$$

Consider two matrices $A, B \in \mathbb{R}^{(m,l)}$. Define

$$A \cdot B := tr(A^\top B)$$
$$\text{where } tr(Z) := \sum_i Z_{ii}$$

- Prove that $A \cdot B$ can be written as an inner product of two vectors as in Equation (1).
- Prove that $A \cdot B = B \cdot A$. It is indeed symmetric even if it does not look like that.
- We know that for every inner product it holds $v \cdot w = \|v\| \|w\| \cos \angle(v, w)$. So what is the cosine angle between $\begin{pmatrix} 1 & -2 \\ -2 & 3 \end{pmatrix}$ and $\begin{pmatrix} -1 & 1 \\ 0 & -2 \end{pmatrix}$? Which angles can constitute the computed cosine angle?

Note: if you know Einstein sum convention, then you can define analogously an inner product between pairs of 3-tensors and any pairs of n-tensors.