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 in class coding: Work with a known $P(x, y)$, Overfitting

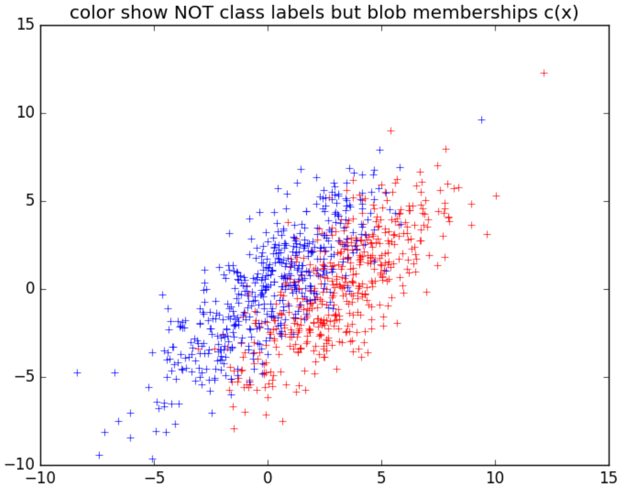
The goal here is two-fold:

- work with an explicit representation of $P(x, y)$
- experience a case of overfitting: a classifier that has very low training error but is not the best one, and performs not so well on evaluation data

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. To do this:
  * draw the random vector $u = (u^{(1)}, u^{(2)})$, where each component $u^{(d)}$ is drawn from a univariate normal distribution with zero mean and variance one.
  * if $C = 1$, then do the transformation:

    $$x = A \cdot u + \mu_1$$

    if $C = 2$, then do the transformation:

    $$x = A \cdot u + \mu_2$$

    where A, μ_i are defined as:

    $$A = \begin{pmatrix} \cos(\pi/4) & +\sin(\pi/4) \\ -\sin(\pi/4) & \cos(\pi/4) \end{pmatrix} \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$$
    $$\mu_1 = (0, 0)$$
    $$\mu_2 = (2.5, 0)$$

    x will be the data sample vector used.
- using the value of C, draw y according to

$$p(y = 0|x, c(x) = 1) = 0.3$$
$$p(y = 0|x, c(x) = 2) = 0.6$$

– print the samples x for $n = 1000$, such that the color of the sample is equal to the value of $c(x)$

– print the samples x for $n = 1000$, such that the color of the sample is equal to the value of y

1.3 Implement a good and a bad classifier.

– draw a dataset D_n with $n = 500$

– Implement a nearest neigbor classifier which is fitted by the data D_n . It works as follows. Suppose we have to label a sample x. The label of $\hat{x}$ will be defined as the label y_* of the sample $(x_*, y_*) \in D_n$ which has nearest euclidean distance between x and x_* among all samples $(x, y) \in D_n$. formally

$$f(\hat{x}) = y_* \text{ such that } (x_*, y_*) = \text{argmin}_{(x,y)\in D_n} \|x - \hat{x}\|$$

– draw a dataset T_k with $k = 1000$. Measure the classification error of the classifier fitted above using D_n on the evaluation/test dataset T_k

– this classifier has obviously classification error 0 on D_n – each data sample from D_n is nearest to itself.

– now use the following classifier on vectors $x = (x^{(1)}, x^{(2)})$:

$$g(x) = \pm\left((+1, +1) \cdot x - 1.25\right)$$
$$f(x) = sgn(g(x)) \text{ must apply the } \pm \text{ to } (+1, -1) \textbf{ and } \text{the bias}$$

Measure the classification error of this classifier on the same evaluation/test dataset T_k from above.

– Why does the 1-nearest neighbor classifier perform worse?

1.4 Data generation explained in depth

- We will draw samples x from a mixture of 2 gaussians. Suppose $c(x) \in \{1, 2\}$ denotes whether sample x was drawn from gaussian 1 or gaussian 2.The densities of the multivariate normal distributions are given as:

$$f(x|c(x) = 1) = \frac{1}{(2\pi)^{d/2} det(\Sigma_1)^{1/2}} \exp(-\frac{1}{2}(x - \mu_1)^T \Sigma_1^{-1}(x - \mu_1))$$
$$f(x|c(x) = 2) = \frac{1}{(2\pi)^{d/2} det(\Sigma_2)^{1/2}} \exp(-\frac{1}{2}(x - \mu_2)^T \Sigma_2^{-1}(x - \mu_2))$$

We assume here equal covariances for both classes with a special shape:

$$\Sigma_1 = \Sigma_2 = \begin{pmatrix} \cos(\pi/4) & +\sin(\pi/4) \\ -\sin(\pi/4) & \cos(\pi/4) \end{pmatrix} \begin{pmatrix} 9 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \cos(\pi/4) & -\sin(\pi/4) \\ +\sin(\pi/4) & \cos(\pi/4) \end{pmatrix}$$

You will see later why I have chosen such a covariance - the main axes of the data cloud are rotated by $\pi/4$ – 45 degrees against the coordinate system.

$R(\alpha) = \begin{pmatrix} \cos(\alpha) & +\sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{pmatrix}$ is a rotation matrix in 2 dimensions (plot 2 dimensional vectors x transformed by it $R(\alpha) \cdot x$ in python, if you do not believe it)

We assume that the means of the two gaussian distributions are different

$$\mu_1 \neq \mu_2$$

- With what probability to draw from gaussian 1 or 2 ? We choose here 50/50, that is our mixture model for samples x is

$$f(x) = f(x|c(x) = 1)0.5 + f(x|c(x) = 2)0.5$$

The meaning of this is: we draw with 50% probability from the normal with $c(x) = 1$ and with the other 50% from the other normal ($c(x) = 2$).
This equation can be derived step by step. In order to avoid confusion,

one needs to see here: x has a density function $f(x)$ such that

$$\int_x f(x)dx = 1$$

, $c(x)$ is a discrete variable with outcomes $c(x) \in \{1, 2\}$ which has a discrete probability such that

$$P(c(x) = 1) + P(c(x) = 2) = 1$$

– the probabilities of all outcomes sum up to one.

So lets derive it step by step. The density $f(x)$ of samples x can be decomposed into 2 disjoint events.
The first event is: x and $c(x) = 1$. The second event is: x and $c(x) = 2$

$$f(x) = f(x, \{c(x) = 1\} \text{ or } \{c(x) = 2\}) = f(x, c(x) = 1) + f(x, c(x) = 2)$$

Recap: $f(x, c(x) = 1)$ means: the probability density of the x and $c(x)$ being equal to 1. This is a joint probability, not a conditional probability. Now lets express it by the conditional probabilities from above:

$$\begin{aligned} f(x) &= f(x, c(x) = 1) + f(x, c(x) = 2) \\ &= f(x|c(x) = 1)P(c(x) = 1) + f(x|c(x) = 2)P(c(x) = 2) \\ &= f(x|c(x) = 1)0.5 + f(x|c(x) = 2)0.5 \end{aligned}$$

because we draw 50/50 from one of the gaussians. We have derived above equation. We see that 0.5 is the probability of observing gaussian blob membership $c(x) = 1$: $P(c(x) = 1) = 0.5$ which makes sense: we draw with 50% chance from blob 1.

- **Recap: How to code the act to draw from one of these gaussians?**

  If we draw for a vector $x = (x^{(1)}, x^{(2)})$ each dimension $x^{(1)}$ independently from a one-dimensional distribution $N(0,1) \sim \frac{1}{(2\pi)^{1/2}} \exp(-\frac{1}{2}x^2)$, then x will be distributed as a two-dimensional (bivariate) gaussian with parameters

  $$\mu = (0,0), \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I_2\,,$$

  where I_2 is the identity matrix for 2 dimensions.

  We need to use a theorem about gaussian distributions and linear mappings:

  Suppose $x = (x^{(1)}, x^{(2)})$ is a 2-dim vector, suppose we do an affine transformation

  $$\begin{aligned} y &= (y^{(1)}, y^{(2)}) = Ax + b \\ A &= \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \\ b &= (b_1, b_2) \end{aligned}$$

  Theorem:

  If $x \sim N((0,0), I_2)$, then the linear transformed $Ax + b$ has normal distribution with parameters $N((b_1, b_2), AA^T)$.

  How to choose A such that $AA^T = \Sigma_1 = \Sigma_2 =$

  $$\ldots = \begin{pmatrix} \cos(\pi/4) & +\sin(\pi/4) \\ -\sin(\pi/4) & \cos(\pi/4) \end{pmatrix} \begin{pmatrix} 9 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \cos(\pi/4) & -\sin(\pi/4) \\ +\sin(\pi/4) & \cos(\pi/4) \end{pmatrix}\ ?$$

  One can see: The rotation matrix on the left is the transpose on the right and:

  $$\begin{pmatrix} 9 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$$

  Therefore:

  $$A = \begin{pmatrix} \cos(\pi/4) & +\sin(\pi/4) \\ -\sin(\pi/4) & \cos(\pi/4) \end{pmatrix} \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$$

  ensures that $AA^T = \Sigma_1 = \Sigma_2$

- what needs to be specified are the means μ_1, μ_2. We use here

  $$\mu_1 = (0,0), \mu_2 = (2.5, 0)$$

- So far we have talked only about x. Now lets specify how to sample y in a pair (x, y).

  The probability of a label y depends on whether the data sample x came from gaussian 1 or gaussian 2. When we draw a sample x then we have at first decided (with 50% chance) whether we draw it from gaussian 1 or gaussian 2. After that we draw x from the gaussian that we have decided for. As a consequence: when we draw x, then we know the value of $c(x)$ – denoting whether x was drawn from gaussian 1 or gaussian 2.
  Therefore: we can use the information from $c(x)$ to define y.

  $p(y = 0|x, c(x) = 1)$ is the probability of label $y = 0$ given that x came from gaussian 1 ($c(x) = 1$). We know that y takes only values in $\{0, 1\}$. Therefore:

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

  Same: $p(y = 0|x, c(x) = 2)$ is the probability of label $y = 0$ given that x came from gaussian 2 ($c(x) = 2$). We define the labels with a probability:

  $$p(y = 0|x, c(x) = 1) = 0.3$$
  $$p(y = 0|x, c(x) = 2) = 0.6$$

- this idea can be extended to more than 2 gaussians obviously

1.5 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:

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

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 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 $p(y = 0|x, c(x) = 1), p(y = 0|x, c(x) = 2)$.
- then plug in the values that you have, use for the homework

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

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

You can start from above equation.