INTRO TO MACHINE LEARNING

classmate
Date
Page

Lecture 1 MIT 6.036: Introduction to Machine Learning (Tamara Broderick)

★ Slide 19 (~~Linear classifiers~~) (Intro)

- You can assume for example, we want to determine ~~whether~~ whether a new born baby will suffer a seizure after being kept in an incubator.

- Assume we have 'n' babies to study.
  For an i^{th} baby, we have two things:
  → Feature vector with 'd' dimensions. These vectors have ~~reasons that the baby~~ parameters such as amount of oxygen in the incubator, temperature of incubator, etc. & they have d values.
  → Label $\in \{-1, 1\}$. This label indicates whether the baby suffer a seizure or not.

★ Slide 90 (Linear Classifiers)

- Hypothesis is a function that for ~~an unknown~~ random new baby with an unknown label; predicts the label from the feature vector.
- Hypothesis class is a class or set of all candidate hypothesis
- In this case (linear classifier), the hypothesis is a line, which divides the dimensional space
- This one (slide 90 example) is obviously a bad hypothesis.

★ Slide 110

- Here, θ & x are two column vectors. & the projection of vector x on vector θ is given by $\frac{\theta^T \cdot x}{|\theta|}$ i.e. dot product of θ & x divided by magnitude of θ.

- This projection can tell how much close are the directions in which θ & x are pointing. You can imagine that if x is rotated more in direction of θ, the projection will increase.

- The projection on θ wont change even if θ is scaled upwards.

★ ~~Slide 119~~ & ~~slide 134~~ Slide 119

~~slide 119~~ & ~~slide 134~~ Slide 119

- Now, the projecting line will always be perpendicular to θ. All the points lying on this line will have same projection on θ. Hence, we can use this line as a classifier.
- All the points above this line will have projection greater than points lying on the line. & all points below this line will have projection lower than points lying on the line.
- We can define the line as :-

  $x : (\theta^T x)/|\theta| = p$

  i.e. set of all points x such that projection of x on θ is p.
- ~~We can alternate~~

~~Slide 134~~ Slide 134

- We can alternatively define the line as:

$x : \theta^T x + \theta_0 = 0$

where, $\theta_0 = -p \cdot |\theta|$

★ Slide 146

- The semicolon in the $h(x; \theta, \theta_0)$ distinguishes x from θ, θ_0 & tells that x is an input, but θ & θ_0 are just identifying parameters.
- Just as we use i to denote i^{th} index or value, we use θ & θ_0 here.

☆ Slide 161 (How good is a linear classifier)

- In the case of loss, ~~we want to~~ it is preferred to differentiate losses from each other. We want to differentiate how much ~~a~~ one type of loss affects us & how other type of loss affects us. (~~We are not talking~~ Here, types of losses does NOT mean 0-1 & assymmetric loss)

- So, we prefer assymetric to 0-1.

- In the case of assymetric, we can compare the $(g=1, a=-1)$ as this: we ~~identified~~ determined using our model for a new baby that this baby will have seizure. So, the doctor checks up on it & ~~it~~ it doesn't have a seizure. So, everything's fine but we got ~~our~~ ~~need~~ loss.

- In case of $(g=-1, a=1)$, i.e. we predict that the baby will ~~have a seizure~~ not have a seizure. So, the doctor doesn't check up on it. But the baby does suffer from seizure & there's no one to tend to it. Now, this is a greater loss than previous one. Hence, we associate higher loss value to it.

★ Learning a classifier (slide Last)

- To answer the last question, the training error of hypothesis returned by alg(Dn; 2) will always be lower than hypothesis returned by alg(Dn; 1).
- You can prove that by proving that the ivarse can't be true. Suppose that alg(Dn, p) spits out hypothesis 'a' & alg(Dn, q) ~~spits out~~ is taken into consideration where q > p. Now, in the 'p' hypothesis, 'a' had the least error. Now when we look for further 'q-p' hypothesis, they may have all hypothesis with error greater than 'a' & will spit out previous result 'a'. If 'q-p' has at least one hypothesis with error less that 'a', it will return that hypothesis. Thus, for ~~p to q~~ q > p, algo(Dn, q) will spit out hypothesis with error strictly lower than or equal to algo(Dn, p).