ECON484 Machine Learning

3. Supervised Learning and Regression

Lecturer:

Bora GÜNGÖREN

bora.gungoren@atilim.edu.tr

Supervised Learning (cnt’d)

- In the basic classification problem we have 2 “classes” (red/blue, sick/not sick, person/not a person, buy/hold, etc.) and we try to learn to make the differential diagnosis from the “supervision” of early examples.
    - Also recall that we had a rough discussion about 2^N.
    - Now it’s time to learn some more theory about supervised learning.

Supervised Learning (cnt’d)

- Say we want to identify the qualities of a particular class, say a “previously (2021) profitable company that will profit next year (2022) as well.”
    - We look at the balance sheets of many companies that were profitable in 2020, and 2021. We also look at the balance sheets of many companies that were profitable in 2020, but not in 2021. So we label them.
    - Then we calculate many financial ratios for both 2020. These are the features we base our learning on.
    - Learning is by labels.

Supervised Learning (cnt’d)

- By learning, we identify the features that are use in the decision, and some arithmetic and/or logical rules to make the decision.
    - We might call the features and the rules a “recognizer.”
    - The companies are the **inputs** but the features are the **input representations**.
    - The rule set we come up with after learning, which will be used to make the decision to label a new input is a **hypothesis class** (ie. x1 > a).
    - The hypothesis that specifically assigns a new input to a class (ie. label it) is a **single hypothesis** (ie. x1 > 0.5) belonging to that class.

Supervised Learning (cnt'd)

- In supervised learning we have a finite number of examples to learn from.
  - From these finite examples, we draw a specific hypothesis and test it with the examples for which we know the answer (ie. label).
  - The **empirical error** is the proportion of errors to the examples. We know these are errors, because we knew the answers.
  - Therefore finding the **best fitting hypothesis** is trying out **all** possible hypothesis from the hypothesis class, and finding the one with the smallest empirical error.

Supervised Learning (cnt’d)

- In supervised learning we have a finite number of examples to learn from.
    - So it becomes a concern, about how many hypothesis can stem from a particular hypothesis class (the **span** of the class).
    - Another concern is the **generality** of each hypothesis. For the rule form x1 > a, if both x1 > 0.5 and x1 > 0.6 end up with the same predictions and same empirical error (with the already available examples), then is the less general (tighter) one better or not?

Supervised Learning (cnt’d)

- In supervised learning we have a finite number of examples to learn from.
  - Assuming the tightness of the hypothesis does not matter, only the results matter.
  - With 2 classes and N examples, we can end up with 2^N different results so we end up with 2^N hypothesis in the hypothesis class.
    - Just out of curiosity, 2^{10} ~ 1.000 and 2^{30} ~ 1 billion. 2^{350} is very large (larger than a googol).
    - More examples will take much more time to evaluate all hypothesis and find the best one.

Supervised Learning (cnt'd)

- The 2^N result is known as Vapnik-Chervonenkis (VC) dimension.
    - The VC terminology is such that any hypothesis that has a different result **shatters** the N points (into classes).
    - VC-dimension teaches us that exhaustive search methods are infeasible for supervised learning.
    - So any time-constrained application result will be **approximately correct** (AC).

Supervised Learning (cnt’d)

- From the computer science perspective, all methods we use for supervised learning are **time-constrained search algorithms**, that try to find a best (so far) hypothesis with the least empirical error.
- However in practice, we have the hypothesis before we test it.
    - From a statistical point of view, we might also **estimate the number of examples necessary to validate this hypothesis**.
    - **Probably Approximate Correct (PAC)** learning focuses on examples from an unknown but fixed distribution, and tries to find the necessary number of examples so that the **empirical error is bound by a number we decide beforehand**.

Supervised Learning (cnt’d)

- PAC gives us a result about how tight (an hypothesis) is OK for a given error rate. Therefore it has practical importance.
- With PAC we can make statements such as:
  - Provided we get ….. examples,
  - we can make predictions with …. confidence probability and with
  - mis-classification error probability at a maximum value of …..
- Noise has the effect of making things difficult, so with noisy data the mis-classification error probability will be large for same number of examples.
  - Increase examples by how much?

Supervised Learning (cnt’d)

- With more classes than 2, we need to understand how larger the problem becomes.
  - With systematic work we can reduce a K-class problem into K separate 2-class problems. So the VC-dimension becomes K times 2^N instead of K^N.
    - Classes A, B, C. First problem A or else.
    - Second problem, given not A, B or else.
    - Third problem given not A, not B, C or else (ie. outlier).

Supervised Learning (cnt’d)

- If we have the means to work in parallel, we can also parallelize our search. However parallel search is not guaranteed to have the same results.
  - Without communication, branch one A or else.
    - Branch one, given not A, B or else.
    - Branch two given not A, C or else.
    - Merge two branches by identifying outliers and classifying both B and Cs.
  - With communication, branch one A or else.
    - Branch one, given not A, given not C from branch two, B or else. (finders keepers).
    - Branch two, given not A, given not B from branch one, C or else.
    - Merge two branches by identifying outliers.

Regression

- In regression out hypothesis is not a classification hypothesis, because there is no Boolean or categorical result.
  - We need to estimate **a continuous function**.
  - We do not know the function, but **we assume its form**. (Hypothesis class).
  - By estimating its parameters, we **end up with a specific function** (hypothesis).
  - We estimate these parameters by trying to **minimize the empirical error** (now calculated differently than a simple rate).
  - Given noise and time-constraints we end up with a PAC result with a **confidence interval**.
- In regression we explain noise simply as **missing variables in the form of the function** we assumed.

Model Selection and Generalization

- In classification problems each new example helps us halve the number of possible hypothesis in our hypothesis class.
    - So after some point (the estimated number of examples) the classification works as required.
    - Adding more examples tightens the hypothesis and ends up as over-fitting.

Model Selection and Generalization

- However, having many examples also gives us the comfort of generating multiple candidates with the required confidence interval based on the examples.
  - So if you have a necessity of N examples, but have N+K examples, then you have 2^K viable alternate hypothesis to choose from.
  - How do we choose? By testing.

Model Selection and Generalization

- Testing works best when you have more than enough examples to learn from.
    - You train multiple alternate models which have the required error rates on the examples (training data).
    - Then you test them on the test data set.
    - Choose the best.
- K-fold cross validation on a sufficiently large data set produces many more alternate models, and many more test data sets.
- Any error introduced by the hypothesis class (ie. form of the rules) is called an **inductive bias**, because the induction comes from the model built by the modeler who has the bias.

Model Selection and Generalization

- An hypothesis class has to have a more complex form (ie. more complex rules and/or more complex functions) in order to span a larger number of hypothesis (or functions).
  - By trying **as simplest form as possible** we will always introduce some inductive bias.
  - Model selection is not the process of eliminating bias, it is a process of **selecting the least harmful bias**.

Model Selection and Generalization

- We **need** a complex hypothesis for a complex problem.
  - If we choose to have a simpler than necessary hypothesis form, we will tend to be be **under-fitting**. This is because a simpler than necessary hypothesis will have more incorrect results.
  - If we choose to have a more complex than necessary hypothesis form, we will tend to be **over-fitting**.
  - The straightforward explanation of noise in regression (missing variables) **inherently assumes that we are always under-fitting** since our model (hypothesis form) is simpler than necessary (because we are missing variables).

Model Selection and Generalization

- There is a three-way trade-off in supervised learning.
    - The complexity of the hypothesis to fit the data (the capacity of the hypothesis class to span a larger number of hypothesis).
    - The amount of training data.
    - The generalization error on the new examples (under-fitting vs over-fitting).

Model Selection and Generalization

- As the complexity increases, the generalization error first decreases and then increases.
- Because we can measure the generalization error, we use this ability to our advantage.
  - We use a validation set (ie. test data set) and measure the generalization error of each model. The less error there is, the less bias we assume.
- The error is also called a **loss function**.
- At this moment we have a big assumption: **All examples are independent and identically distributed (IID) although we may not know which distribution.**

Questions?

CONTACT:
bora.gungoren@atilim.edu.tr