Classifiers

CS 5703: Machine Learning Practice

Classifiers

Given some input, which of several categories does this situation belong?

- Number of categories (classes) is finite
- Used in many types of problems:
  - Is the input image an example of a cat, dog, horse?
  - Is this loan a good risk?
  - Is the tumor malignant or benign?
  - Is that a stop sign or a speed limit sign? (or others)

Classifier Formulation

- In the general case, input data can be numerical or categorical
- For our first set of examples, we will assume numerical
  - And: categorical data can be transformed into numerical using One-Hot-Encoding
- For now, we will also assume two classes

Classifier Formulation

- With N-dimensional numerical data, training samples are labeled points (labels correspond to the classes)
- Task: identify a N-1 dimensional surface that separates the points in a way that respects the labels

- When N=2, the surface becomes a curve
  - And: the simplest (interesting) curve is a line

Measuring Classifier Performance

Measuring Classifier Performance

One straw-man possibility for measuring the performance of a specific classifier: count the number of training examples that are labeled incorrectly by the current parameters

Measuring Classifier Performance

One straw-man possibility for measuring the performance of a specific classifier: count the number of training examples that are labeled incorrectly by the current parameters

- Many solutions look the same by this metric
- For this metric, it is not clear how to change the parameters so we can improve the classifier

A First Classifier Learning Algorithm

- Randomly choose parameters
- Measure error
- While error is too large:
  - Make small random choices to the parameters
  - Measure error
  - If the error does not become larger, then keep the new parameters
- Done

A First Classifier Learning Algorithm

This is easy to implement, but:

- We could go many random steps before improving performance
- We will randomly choose a solution that minimizes cost
  - But, not all of these solutions are really the same

Logistic Regression

CS/DSA 5703: Machine Learning Practice

Logistic Regression

Motivation

- Want to have a smooth relationship between parameters and the cost
  - I.E., we want the cost function to be differentiable with respect to the parameters
- Want to acknowledge that examples near the dividing line are still not really acceptable
  - Instead, we want all samples far away from the dividing line

Logistic Regression

Approach: add a non-linearity onto the function

- Dividing curve is still a line
- But, we can use a different cost function that is smooth in the parameter space

New Algorithm: Stochastic Gradient Descent

- Randomly choose parameters
- Measure error
- While error is too large:
  - For one or more training samples: compute the derivative of error with respect the parameters (i.e., the slope):

    For each i, compute: $\frac{\partial E}{\partial w_i}$

  - Change the parameters in the opposite direction

    For each i: $w_i \leftarrow w_i - \alpha \frac{\partial E}{\partial w_i}$

- Done

New Algorithm: Stochastic Gradient Descent

Notes:

- Stochastic aspect: we only compute the cost with respect to one or a small number of training samples
  - Often this is a sufficient estimate of the gradient
- Computation of the gradient is straight forward
- Depending on the training set, error may always be large
  - Change of algorithm: loop until error stops changing

Classes in the Infant Kinematic Data

CS 5703: Machine Learning Practice

Infant-Robot Interaction

The robot responds to the infant for a variety of reasons, including:

- **Power steering**: if the infant pushes against the ground with a large enough force in a given direction, then the robot responds with a large movement in the forward, left or right directions
- **Gesture-Based Control**: if the infant moves their arms or legs in one of several specific ways, the robot also responds with a large movement in the forward, left or right directions

Example: Infant Kinematic Data

New columns in our infant kinematic data:

- Positions of more than just the wrists
- Assistance action type being given to the infant:
  - 0 = none
  - 1 = forward (power steering)
  - 2 = backward
  - 3 = left
  - 4 = right
  - 5 = forward (gesture)
  - 6 = backward
  - 7 = left
  - 8 = right

Preprocessing

- Compute velocity for all kinematic columns
- Drop all samples with NaNs

First Prediction Problem

Given position and velocity of all points on the body (wrists, shoulders, knees, ankles, toes): predict whether the robot is currently providing assistance

- Can be power steering or gesture-based (action type > 0)

Example: First Behavior Classifier

Stochastic Gradient Descent Classifier

- Provides a variety of linear-based classifiers
- Allows us to select from a range of different loss metrics
  - loss = ‘log_loss’ selects logistic regression

Let’s Build a Model with SGD…

…

Classifier Performance Measures

CS 5703: Machine Learning Practice

Learned Model

So far:

- Model computes a score for a given input that ranges between 0 and 1 (due to the sigmoid)
- If the score is larger than some threshold, then we say that the model has detected a positive example
  - We call this output “+”
  - For other cases, we call the output “-”
  - For logistic regression, this default threshold is 0.5

Basis of our Performance Measures

- **True Positives (TP)**: the examples for which the model correctly answered “+”
- **True Negatives (TN)**: the examples for which the model correctly answered “-”
- **False Positives (FP)**: the examples for which the model incorrectly answered “+”
- **False Negatives (FN)**: the examples for which the model incorrectly answered “-”

...

Multi-Class Classification

CS 5703: Machine Learning Practice

Multi-Class Classification

- A linear decision surface (such as what is used in SGDClassifier) is necessarily binary
- To address multiple classes, we must construct a set of binary classifiers
  - Predictions over this set are combined together to create a single, monolithic prediction for each input

Multi-Class Classification

One-versus-one approach:

- For every pair of classes, create a classifier that distinguishes examples from the two classes
- We assume that the two classifiers randomly assign a label to all other example types (not necessarily a good assumption)
- Need N^2 classifiers

Multi-Class Classification

One-versus-all approach:

- For each class, create a classifier that distinguishes examples from one class and all other classes
- Need N classifiers
- Decision surfaces can be complex, which are hard to model with a linear surface

Multi-Class Classification with the SGDClassifier

- SGDClassifier automatically detects when it is faced with a multi-class situation
- Unless forced, it will choose oneVone or oneVall, depending on the number of classes

Multi-Class Classification with the SGDClassifier

Example :

- 3 classes: gesture forward, gesture left/right, all others
- Construct model, examine predictions, confusion matrix and class probabilities

Example II:

- Same, but with cross-validation

Multi-Class Classification with the SGDClassifier

Example III:

- RandomForestClassifier

Final Notes

This particular classification problem is a challenge:

- Example uses only a small amount of data
- Labeling process leaves a lot to be desired
  - Only labeling movement as positive
  - But, one sample before the positive label will have very similar positions and velocities (and yet be labeled as negative)
  - In practice: we tend to censor these nearby samples

Classifier Summary

CS 5703: Machine Learning Practice

Classifiers

SGDClassifier

- Numerical data
- Limited to constructing linear decision surfaces
- Must take extra steps to address multi-class cases

SGDClassifier Parameters

Some key parameters:

- Loss function
- Regularization (L1, L2 or both)
- Maximum number of iterations
- Tolerance
- Learning rate (and is it constant or adaptive)
- Early stopping (using a validation data set)

Classifiers

Looking forward to other types of classifiers:

- Non-linear decision surfaces
- Picking decision surfaces as conservatively as possible
- Allowing the algorithm to choose some training samples to ignore
- Categorical data as inputs

Classifier Metrics

- Precision & recall
- True positive rate & true negative rate
- Receiver Operator Characteristic Curve
  - Area under the ROC Curve (AUC)
- Skill scores
  - We looked at Pierce Skill Score (PSS), but there are others that address different properties

Cross-Validation

- Only report performance for data that are not used to select model parameters
- Cross-Validation explicitly does this in situations where data samples are hard to come by

More on this topic later in the semester…