ECE M146 Introduction to Machine Learning

Prof. Lara Dolecek

ECE Department, UCLA

Today’s Lecture

Recap:

- Perceptron and linear least squares; gradient descent

New topics:

- Logistic regression
- More on loss functions

Today’s Lecture

Recap:

- Perceptron and linear regression; gradient descent

New topics:

- Logistic regression
- More on loss functions

Recap

- Perceptron is on-line algorithm for binary classification; at test time, outputs one of two choices

- Linear least squares for linear regression; at test time, outputs a real-valued number

- (Stochastic) gradient descent can be used for both.

Today's Lecture

Recap:

- Perceptron and linear regression; gradient descent

New topics:

- Logistic regression
- More on loss functions

Logistic regression

- Can be viewed as an in-between method in the sense that it outputs a real value, but it is used for classification.

- Like the methods studied so far, it represents a linear model.

- It is an instance of a probabilistic discriminative model, where we model $p(y|x)$. As such it is modeling-wise more complex than those e.g., perceptron that are described by a discriminant function.
- Later on, we will see generative models such as naïve Bayes, that model $p(x|y)$.

Logistic regression – key idea

- Probabilistically capture the confidence of the classified point.

- The closer the point is to the decision boundary, the less confidence it has in its value; the further away the point is from the decision boundary, the more confidence it has in its value.

- Contrast with perceptron.

Logistic sigmoid

- A convenient function that we will use in logistic regression.
- Picture:
- Math:
- Evaluations:

Inference set-up

- Label is again binary, but we switch to {0,1} for mathematical convenience.
- This does not make any difference conceptually as it is still binary classification.
- Define the following functions:

- These functions are unknown. Our goal is to model these conditional probabilities.

Logistic regression modeling

- In logistic regression, we use the following modeling for the conditional probabilities:

- Again, as in perceptron and in linear least squares, the goal is to find the best vector w under an appropriately chosen loss function.

A useful property of the logistic sigmoid

Intuition on why this functional format

- Consider the conditional probability $P(y=1|x)$.

Maximization of the likelihood

- Since we now focus on $P(y|x)$, the goal is to maximize the following:

- We have N data points.
- How ?

- We now have a function to minimize.

Optimization details

- Unfortunately, we cannot take the derivatives as in LLS, but we can apply gradient descent!
- Recall the expressions for the conditional probabilities:
- Mathematical trick:
- Check:

Back to our set up

- Then, and this is why we applied log, bring the exponents down to get:

- Take the gradient with respect to w.

First, an auxiliary result

- A result from matrix calculus we'll need:

- Because:

Back to our minimization problem

Back to our minimization problem – ctd.

- Ctd.

- This is the gradient.

Connection to cross-entropy

- Consider two RVs, X and Y. Suppose X is distributed as a Bernoulli RV with parameter p and Y is distributed as a Bernoulli RV with parameter q.

- Cross-entropy is defined as:

- Note that the loss is a scaled version of the cross-entropy.

Connections to quadratic loss

- Recall that in linear regression with quadratic loss we saw:

- Derivatives were of the same format:

Further discussion on logistic regression

- Mathematical derivations thus far were for the batch gradient descent; this method can also be done with stochastic gradient descent using one data point at the time.

- When data is linearly separable, this method can result in severe overfitting:

- All probabilities degenerate to 1.

At testing time

- Once we have the weight vector, at testing time, we perform the following.

Today’s Lecture

Recap:

- Perceptron and linear regression; gradient descent

New topics:

- Logistic regression
- More on loss functions

Loss functions

- There are different loss functions, and we have already seen some. Consider:
- 0/1 loss
- Squared loss:
- Logistic loss:
- Exponential loss:
- Hinge loss:

- They differ in the kind of penalty they incur.

Loss functions

- Hinge loss – used in SVM

- Logistic regression uses cross-entropy loss, which is equivalent to logistic loss, up to scaling.