Optimization in Statistics

Owen Ward

February 15, 2019

What is optimization?

- Have actually already seen several examples of optimization in this class.
- In statistics we try to fit models which approximate our data.
- Want to choose the model which “best” approximates our data.
- Want to “optimize” over a selection of models to find the “best” one.

Linear Regression

- In this setting we have the model

$$y = \alpha + \beta x,$$

and we wanted to find the best pair $\hat{\alpha}, \hat{\beta}$ which best fit our data. This is a simple example of optimization.

A quick review of Calculus

- If we differentiate a function and find the values where this derivative are zero, these are turning points of the function.
- We establish if these are local maxima or local minima by evaluating the second derivative at these turning points.
- For certain types of functions (convex/concave), then these local optima may be global.

Finding the max/min of a function

- Write this down
- When fitting a model, we want to come up with a function which describes the difference between our data and the model, and minimize this function.
- This is often known as the loss function.
- Some examples of loss functions are:

Global Max

- Certain types of functions have only one maximum, and it can be found in a straightforward way, using . . .

Newton-Rhapson method

- Want to find roots of some function $f(x)$.
- Start with some inital estimate x_0
- Improve this estimate iteratively with the formula,

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

  until it changes only by a small amount.
- Once you start suitably close to the zero you will reach it.

More general methods

- Gradient descent/etc

Functions without a unique maximum

- Many common functions are more complex and do not have a global maximum, instead several (or possibly infinite) local maximums.
- Can be difficult (or impossible) to find which of these maximums is the global maximum.
- When optimizing high dimensional functions this can be challenging.

Methods to attempt to find the global max

- How could we try get around this?

Methods to attempt to find the global max

- How could we try get around this?
- One way is to start some of the methods described above from different (maybe random) locations, and find the overall maximum of the maximum each start finds.

Methods to attempt to find the global max

- How could we try get around this?
- One way is to start some of the methods described above from different (maybe random) locations, and find the overall maximum of the maximum each start finds.
- No guarantee this will work, but as we will see, is often all that can be used.

Clustering

- Clustering is an extremely common method in statistics and data science.
- It is an unsupervised learning problem. Given some data, we want to try find clusters in the data which reveal interesting relationships.
- This is different to classification, where there are some known labels and we want to predict these labels for some new data.

K-means

- K-means clustering is one of the most common clustering methods.
- It aims to partition data into k groups. Each of the k groups has a center, and a data point is assigned to the cluster corresponding to the nearest center.
- The algorithm tries to find centers which create clusters which are close together, and classifies points to the corresponding closest center.

K-means

- To optimise this, for a fixed k, we want to minimize the distance from each data point and the center of the cluster it is classified to.

K-means

- To optimise this, for a fixed k, we want to minimize the distance from each data point and the center of the cluster it is classified to.
- Mathematically, given that we break the data into k clusters $S_1, \dots, S_k$, each with mean $\mu_1, \dots, \mu_k$, we want to minimize

$$\sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2.$$

K-means

- To optimise this, for a fixed k, we want to minimize the distance from each data point and the center of the cluster it is classified to.
- Mathematically, given that we break the data into k clusters $S_1, \ldots, S_k$, each with mean $\mu_1, \ldots, \mu_k$, we want to minimize

$$\sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2.$$

- This looks tricky to solve, and it looks like it might not have a global maximum.

K-means

- Thinking about how we might try find a maximum of this, we need to maximise two things at the same time: The centers and which cluster we assign each point to.

K-means

- Thinking about how we might try find a maximum of this, we need to maximise two things at the same time: The centers and which cluster we assign each point to.
- Updating the centers might change which cluster we would assign a point to.

K-means

- Thinking about how we might try find a maximum of this, we need to maximise two things at the same time: The centers and which cluster we assign each point to.
- Updating the centers might change which cluster we would assign a point to.
- Updating which cluster we assign points too might change the best center for some or all groups.

K-means

- Thinking about how we might try find a maximum of this, we need to maximise two things at the same time: The centers and which cluster we assign each point to.
- Updating the centers might change which cluster we would assign a point to.
- Updating which cluster we assign points too might change the best center for some or all groups.
- For large amounts of data this requires computing distances many times, which can be difficult for high dimensional data also.

K-means

- Thinking about how we might try find a maximum of this, we need to maximise two things at the same time: The centers and which cluster we assign each point to.
- Updating the centers might change which cluster we would assign a point to.
- Updating which cluster we assign points too might change the best center for some or all groups.
- For large amounts of data this requires computing distances many times, which can be difficult for high dimensional data also.
- How could one go about doing this?

K-means

- Thankfully, the natural way of optimising this works well in practice.

K-means

- Thankfully, the natural way of optimising this works well in practice.
- Update the cluster centers, then update the cluster each point is placed in.

K-means

- Thankfully, the natural way of optimising this works well in practice.
- Update the cluster centers, then update the cluster each point is placed in.
- Then repeat this many times until the clusters stop changing.

K-means

- Thankfully, the natural way of optimising this works well in practice.
- Update the cluster centers, then update the cluster each point is placed in.
- Then repeat this many times until the clusters stop changing.
- There is generally no theoretical reason for this to work but does in practice.

An example

Other optimization methods

- There are lots of more advanced methods to optimize functions.
- To be brief, they all do gradient descent, or some slight variant of it.