10 Data Science Questions

Victor Lu

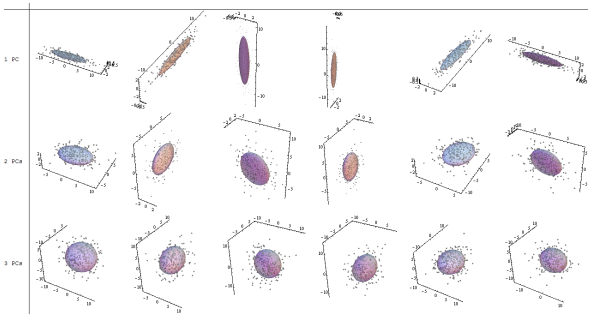
November 29, 2020

Q1

Explain PCA.

Q1

Explain PCA.



Q2

What is a kernel?

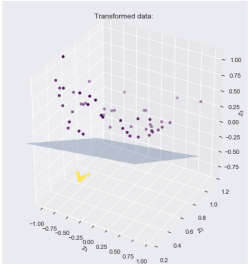
Q2

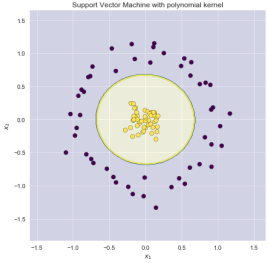
What is a kernel?

Suppose we have a mapping $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ that brings our vectors in $\mathbb{R}^n$ to some feature space $\mathbb{R}^m$. Then the dot product of x and y in this space is $\varphi(\mathrm{x})^T \varphi(\mathrm{y})$. A kernel is a function k that corresponds to this dot product, i.e. $k(\mathrm{x}, \mathrm{y}) = \varphi(\mathrm{x})^T \varphi(\mathrm{y})$.

Why is this useful? Kernels give a way to compute dot products in some feature space without even knowing what this space is and what is φ.

For example, consider a simple polynomial kernel $k(\mathrm{x}, \mathrm{y}) = (1 + \mathrm{x}^T \mathrm{y})^2$ with $\mathrm{x}, \mathrm{y} \in \mathbb{R}^2$. This doesn't seem to correspond to any mapping function φ, it's just a function that returns a real number.





Q3

How does Support Vector Machine (SVM) work?

Q3

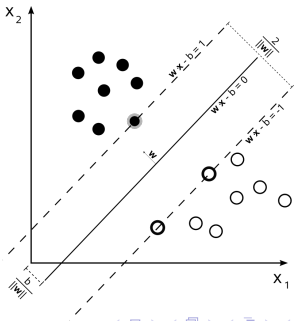
How does Support Vector Machine (SVM) work?

The SVM aims at satisfying two requirements:
1. The SVM should maximize the distance between the two decision boundaries. We want to maximize the distance between the hyperplane defined by $\mathrm{w}^{\mathsf{T}}\mathrm{x} + b = -1$ and the hyperplane defined by $\mathrm{w}^{\mathsf{T}}\mathrm{x} + b = 1$. This distance is equal to $\frac{2}{\|\mathrm{w}\|}$. We want to solve $\max_{\mathrm{w}} \frac{2}{\|\mathrm{w}\|}$.

2. The SVM should also correctly classify all $\mathrm{x}^{(i)}$, which means $y^{(i)}(\mathrm{w}^{\mathsf{T}}\mathrm{x}^{(i)} + b) \geq 1, \forall i \in \{1, \ldots, N\}$

Which leads us to the following quadratic optimization problem:

$$\min_{\mathrm{w}, b} \quad \frac{\|\mathrm{w}\|}{2},$$
$$s.t. \quad y^{(i)}(\mathrm{w}^{\mathsf{T}}\mathrm{x}^{(i)} + b) \geq 1 \quad \forall i \in \{1, \ldots, N\}$$



Q4

What is the influence of C in SVMs?

Q4

What is the influence of C in SVMs?

One can relax the constraints of the hard-margin SVM by introducing so-called slack variables $\xi^{(i)}$. Note that each sample of the training set has its own slack variable. This gives us the following quadratic optimization problem:

$$\begin{aligned} \min_{\mathrm{w},b} \quad & \frac{\|\mathrm{w}\|}{2} + C\sum_{i=1}^{N}\xi^{(i)}, \\ s.t. \quad & y^{(i)}(\mathrm{w}^{\mathsf{T}}\mathrm{x}^{(i)} + b) \geq 1 - \xi^{(i)}, \quad \forall i \in \{1,\ldots,N\} \\ & \xi^{(i)} \geq 0, \qquad\qquad\qquad\qquad\quad \forall i \in \{1,\ldots,N\} \end{aligned}$$

This is the soft-margin SVM. C is a hyperparameter called penalty of the error term.

Q5

What is the difference between test set and validation set?

Q5

What is the difference between test set and validation set?

Training set: a set of examples used for learning: to fit the parameters of the classifier In the Multilayer Perceptron (MLP) case, we would use the training set to find the "optimal" weights with the back-prop rule

Validation set: a set of examples used to tune the parameters of a classifier In the MLP case, we would use the validation set to find the "optimal" number of hidden units or determine a stopping point for the back-propagation algorithm

Test set: a set of examples used only to assess the performance of a fully-trained classifier In the MLP case, we would use the test to estimate the error rate after we have chosen the final model (MLP size and actual weights) After assessing the final model on the test set, YOU MUST NOT tune the model any further!

Why separate test and validation sets? The error rate estimate of the final model on validation data will be biased (smaller than the true error rate) since the validation set is used to select the final model After assessing the final model on the test set, YOU MUST NOT tune the model any further!

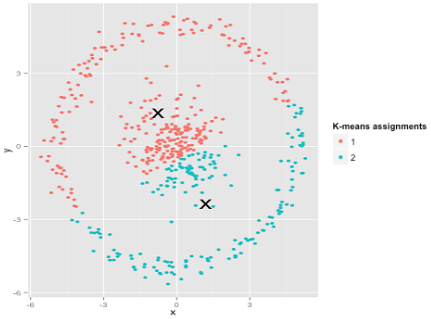
Q6

What are the drawbacks of K-means?

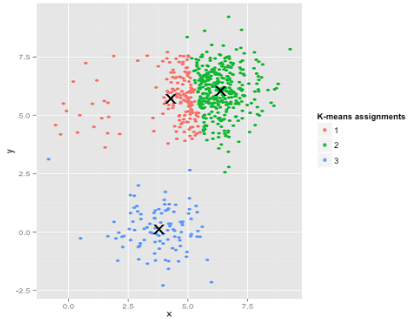
Q6

What are the drawbacks of K-means?

Non-Spherical Data:



Unevenly Sized Clusters:



Q7

Why bootstrapping works: if we are resampling from our sample, how is it that we are learning something about the population rather than only about the sample?

Q7

Why bootstrapping works: if we are resampling from our sample, how is it that we are learning something about the population rather than only about the sample?

You can either:

- make some assumptions about the shape of the population (Normal, or Bernoulli, etc.)
- or you can use the information in the sample you actually have

If you take the sample you have and sample from it. The sample you have is also a population, just a very small discrete one; it looks like the histogram of your data. Sampling 'with replacement' is just a convenient way to treat the sample like it's a population and to sample from it in a way that reflects its shape.

This is a reasonable thing to do because not only is the sample you have the best, indeed the only information you have about what the population actually looks like, but also because most samples will, if they're randomly chosen, look quite like the population they came from. Consequently it is likely that yours does too.

Resampling is not done to provide an estimate of the population distribution–we take our sample itself as a model of the population. Rather, resampling is done to provide an estimate of the sampling distribution of the sample statistic in question.

Q8

When conducting multiple regression, when should you center your predictor variables?

Q8

When conducting multiple regression, when should you center your predictor variables?

Centering/scaling does not affect your statistical inference in regression models - the estimates are adjusted appropriately and the p-values will be the same.

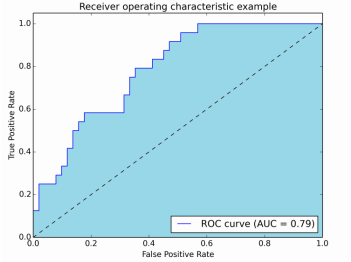
when you're trying to sum or average variables that are on different scales, perhaps to create a composite score of some kind. Without scaling, it may be the case that one variable has a larger impact on the sum due purely to its scale, which may be undesirable.

Q9

What is AUC?

Q9

What is AUC?



- True positive rate (TPR), aka. sensitivity, hit rate, and recall, which is defined as $\frac{TP}{TP+FN}$.
- False positive rate (FPR) , aka. fall-out, which is defined as $\frac{FP}{FP+TN}$.

To combine the FPR and the TPR into one single metric, we first compute the two former metrics with many different threshold (for example $0.00; 0.01, 0.02, \ldots, 1.00$) for the classification model, then plot them on a single graph, with the FPR values on the x-axis and the TPR values on the y-axis. The resulting curve is called ROC curve, and the metric we consider is the AUC of this curve, which we call AUROC.

Q10

What is batch size in a neural network?

Q10

What is batch size in a neural network?

- one epoch = one forward pass and one backward pass of all the training examples
- batch size = the number of training examples in one forward/backward pass. The higher the batch size, the more memory space you'll need.
- number of iterations = number of passes, each pass using batch size number of examples. To be clear, one pass = one forward pass + one backward pass (we do not count the forward pass and backward pass as two different passes).

Example: if you have 1000 training examples, and your batch size is 500, then it will take 2 iterations to complete 1 epoch.