Latest Version: 6.0

Question: 1

Feature Hashing approach is "SGD-based classifiers avoid the need to predetermine vector size by simply picking a reasonable size and shoehorning the training data into vectors of that size" now with large vectors or with multiple locations per feature in Feature hashing?

A. Is a problem with accuracy
B. It is hard to understand what classifier is doing
C. It is easy to understand what classifier is doing
D. Is a problem with accuracy as well as hard to understand what classifier us doing

Answer: B

Explanation:
FEATURE HASHING
SGD-based classifiers avoid the need to predetermine vector size by simply picking a reasonable size and shoehorning the training data into vectors of that size. This approach is known as feature hashing. The shoehorning is done by picking one or more locations by using a hash of the name of the variable for continuous variables or a hash of the variable name and the category name or word for categorical, textlike, or word-like data.
This hashed feature approach has the distinct advantage of requiring less memory and one less pass through the training data, but it can make it much harder to reverse engineer vectors to determine which original feature mapped to a vector location. This is because multiple features may hash to the same location. With large vectors or with multiple locations per feature, this isn't a problem for accuracy but it can make it hard to understand what a classifier is doing.
An additional benefit of feature hashing is that the unknown and unbounded vocabularies typical of word-like variables aren't a problem.

Question: 2

What are the advantages of the Hashing Features?

A. Requires the less memory
B. Less pass through the training data
C. Easily reverse engineer vectors to determine which original feature mapped to a vector location

Answer: AB

Explanation:
SGD-based classifiers avoid the need to predetermine vector size by simply picking a reasonable size and shoehorning the training data into vectors of that size. This approach is known as feature hashing. The

shoehorning is done by picking one or more locations by using a hash of the name of the variable for continuous variables or a hash of the variable name and the category name or word for categorical, textlike, or word-like data.

This hashed feature approach has the distinct advantage of requiring less memory and one less pass through the training data, but it can make it much harder to reverse engineer vectors to determine which original feature mapped to a vector location. This is because multiple features may hash to the same location. With large vectors or with multiple locations per feature, this isn't a problem for accuracy but it can make it hard to understand what a classifier is doing.

An additional benefit of feature hashing is that the unknown and unbounded vocabularies typical of word-like variables aren't a problem.

Question: 3

Question-3: In machine learning, feature hashing, also known as the hashing trick (by analogy to the kernel trick), is a fast and space-efficient way of vectorizing features (such as the words in a language), i.e., turning arbitrary features into indices in a vector or matrix. It works by applying a hash function to the features and using their hash values modulo the number of features as indices directly, rather than looking the indices up in an associative array. So what is the primary reason of the hashing trick for building classifiers?

A. It creates the smaller models
B. It requires the lesser memory to store the coefficients for the model
C. It reduces the non-significant features e.g. punctuations
D. Noisy features are removed

Answer: B

Explanation:

This hashed feature approach has the distinct advantage of requiring less memory and one less pass through the training data, but it can make it much harder to reverse engineer vectors to determine which original feature mapped to a vector location. This is because multiple features may hash to the same location. With large vectors or with multiple locations per feature, this isn't a problem for accuracy but it can make it hard to understand what a classifier is doing.

Models always have a coefficient per feature, which are stored in memory during model building. The hashing trick collapses a high number of features to a small number which reduces the number of coefficients and thus memory requirements. Noisy features are not removed; they are combined with other features and so still have an impact.

The validity of this approach depends a lot on the nature of the features and problem domain; knowledge of the domain is important to understand whether it is applicable or will likely produce poor results. While hashing features may produce a smaller model, it will be one built from odd combinations of real-world features, and so will be harder to interpret.

An additional benefit of feature hashing is that the unknown and unbounded vocabularies typical of word-like variables aren't a problem.

Question: 4

Suppose A, B , and C are events. The probability of A given B , relative to P(|C), is the same as the probability of A given B and C (relative to P). That is,

A. P(A,B|C) P(B|C) =P(A|B,C)
B. P(A,B|C) P(B|C) =P(B|A,C)
C. P(A,B|C) P(B|C) =P(C|B,C)
D. P(A,B|C) P(B|C) =P(A|C,B)

Answer: A

Explanation:
From the definition, P(A,B|C) P(B|C) =P(A,B.C)/P(C) P(B.C)/P(C) =P(A,B.C)
P(B,C) =P(A|BC)
This follows from the definition of conditional probability, applied twice: P(A,B)=(PA|B)P(B)

Question: 5

What is the considerable difference between L1 and L2 regularization?

A. L1 regularization has more accuracy of the resulting model
B. Size of the model can be much smaller in L1 regularization than that produced by L2-regularization
C. L2-regularization can be of vital importance when the application is deployed in resource-tight environments such as cell-phones.
D. All of the above are correct

Answer: B

Explanation:
The two most common regularization methods are called L1 and L2 regularization. L1 regularization penalizes the weight vector for its L1-norm (i.e. the sum of the absolute values of the weights), whereas L2 regularization uses its L2-norm. There is usually not a considerable difference between the two methods in terms of the accuracy of the resulting model (Gao et al 2007), but L1 regularization has a significant advantage in practice. Because many of the weights of the features become zero as a result of L1-regularized training, the size of the model can be much smaller than that produced by L2-regularization. Compact models require less space on memory and storage, and enable the application to start up quickly. These merits can be of vital importance when the application is deployed in resourcetight
environments such as cell-phones.
Regularization works by adding the penalty associated with the coefficient values to the error of the hypothesis. This way, an accurate hypothesis with unlikely coefficients would be penalized whila a somewhat less accurate but more conservative hypothesis with low coefficients would not be penalized as much. 81