

## Machine Learning Crash Course – Week 2 Quiz Q&A

**Q1- Which of the above options are true about logistic regression?**

- I- Logistic regression is a linear model
- II- Logistic regression is a non-linear model
- III- Logistic regression generates a probability—a value between 0 and 1
- IV- Mean Squared Error is the loss function for logistic regression

- I-IV
- I-III ✓
- III-IV
- II-IV
- II-III

**A1-** Logistic Regresyonda çıktılar girdinin ve parametrelerin toplamlarına bağlı olduğu için lineer bir modeldir ve  $1/(1 + e^{(-z)})$  fonksiyonu 0 ile 1 arasında değerler ürettiğinden lojistik regresyon olasılık değerleri verir. Lojistik regresyonun loss fonksiyonu Log-Loss fonksiyonudur. Dolayısıyla cevap I ve III ifadelerini içeren B şıkkıdır.

**Q2- Which of the following statements is wrong?**

- The loss function for linear regression is squared loss.
- The loss function for logistic regression is Log Loss.
- Without regularization, the asymptotic nature of logistic regression would keep driving loss towards 0 in high dimensions.
- If  $x$  represents a number, then  $\text{sigmoid}(x)$  will yield a value between -1 and 1. ✓

**A2-** Sigmoid fonksiyonunun çıktısı 0 ile 1 arasında bir sayı olmalıdır, dolayısıyla yanlış olan ifade son seçenektir.

**Q3- When is using sparse representation efficient in the Machine Learning model?**

- Handle and remove outliers
- To transform highly skewed variable into a more normalized dataset to avoid overfitting
- To reduce and scale high computing resource usage where large of feature binary vector have 0 values ✓
- To normalize features that have very different range to improves the numerical stability of the model and often reduces training time.

**A3-** Sparse representation (seyrek, aralıklı gösterim), yalnızca sıfır olmayan öğeleri saklayan bir tensörün temsildir. Kategorik verilerimizin çok fazla değerden oluştuğunu varsayalım (100.000 farklı değer) Bunları ikili vektör temsiline dönüştürdüğümüzde, değerlerimizin büyük kısmı 0 olacaktır, bu da sparse (seyrek, aralıklı) olduğu anlamına gelir. Eğer ikili vektör temsilini bu şekilde belirtirsek, bu vektörler işlenirken hem depolama hem de hesaplama süresi açısından verimsiz temsil olacaktır. B seçeneği için log transformation (log dönüşümü) kullanılabilir. D seçeneği için de min-max normalizasyonu, z-puanlı normalizasyon gibi teknikler kullanılabilir.

**Q4- Which of the following statement(s) about regularization are/is true?**

- I- Using too large a value of the  $\lambda$  can cause your hypothesis to underfit the data
  - II- Using too large a value of the  $\lambda$  can cause your hypothesis to overfit the data; this can be avoided by reducing  $\lambda$
  - III- Because regularization causes loss to no longer be convex, gradient descent may not always converge to the global minimum (when  $\lambda > 0$ , and when using an appropriate learning rate  $\alpha$ )
  - IV- Using a very large value of  $\lambda$  cannot hurt the performance of your hypothesis; the only reason we do not set  $\lambda$  to be too large is to avoid numerical problems
- II, IV
  - I, III
  - I ✓
  - III

**A4-** Yanıt yalnızca I olan C seçeneğidir.

I- $\lambda$ 'nın büyük değeri, büyük bir düzenlenme cezası (regularization penalty) ile sonuçlanır. Böylelikle datayı underfit yapan basit modeller için güçlü bir öncelik olmuş olur.

II- Çok büyük  $\lambda$  değeri, eğitilen veri setinin (training data) overfit değil underfit olmasına yol açar.

III- Düzenli lojistik regresyon (regularized logistic regression) ve düzenli doğrusal regresyonun (regularized linear regression) her ikisi de dışbükeydir(convex) ve bu nedenle gradyan inişi (gradient descent) hala minimum (global min) seviyeye yaklaşacaktır.

IV- Çok büyük  $\lambda$  değeri, eğitilen veri setinin (training data) overfit değil underfit olmasına yol açar.

**Q5- Which of the following statements are true about one-hot encoding?**

- I- One-hot encoding can be used to numeric data that you do not want to directly multiply by a weight, such as serial number of a product.
- II- The length of the one-hot encode vector is equal to number of observations in the data.
- III- One-hot encoding represents the feature as integers like 0,1,2,3 etc.
- IV- Since models cannot learn from categorical data, we are using one-hot encoding in machine learning problems.

- I – III
- II – IV
- III – IV
- I – IV ✓
- I – II

**A5-** Doğru yanıt I-IV olacaktır.

- I. One-hot encoding seri numarası, kimlik numarası, posta kodu gibi nümerik gösterimde olan ancak sayısal anlamda bir weight ile çarpıldığında anlam ifade etmeyen özellikler için kullanılabilir.
- II. One-hot encoding ile üretilen vektörün uzunluğu datamızın uzunluğuna değil, data içerisindeki benzersiz (unique) özelliklerin sayısına eşittir.

- III. One-hot encoding ilgili satırdaki ilgili özellik mevcutsa 1 değilse 0 şeklinde gösterim yapar, kaç kere geçtiğini saymaz. Boolean yani True/False gibi çalışır.
- IV. Kategorik datalar modele input olarak verilemez. Bunun sebebi string gösterime sahip bir özelliği weight ile çarpıp bir sonuç elde edemememizdir. Şehir adı, sokak adı gibi bilgiler modele one-hot encoding kullanılarak verilebilir.

**Q6- Which of the following statement about the feature cross is false?**

- A feature cross is a synthetic feature that encodes non-linearity in the feature space by multiplying two or more input features together.
- A feature cross might be formed by multiplying the values of two or more features.
- A feature cross cannot be formed by squaring a single feature. ✓
- Linear learners scale well to massive data. Using feature crosses on massive data sets is one efficient strategy for learning highly complex models.

**A6-** Modele doğrusal olmayan ilişkiler eklemek için bazı durumlarda tek bir özelliğin karesini alarak özellik çaprazlama (feature cross) özelliğini kullanabiliriz. Yani C yanlıştır.

**Q7- In regularization, the model can be defined as follow. In this case, what is the purpose of lambda?**

$\text{minimize}(\text{Loss}(\text{Data}|\text{Model}) + \lambda \text{ complexity}(\text{Model}))$

- Lambda is used for simplicity; therefore, loss is minimized.
- If lambda value is too high, the model will be more complex, and it is a risk of overfitting data.
- If lambda value is too low, the model will be more simpler and it is a risk of underfitting data.
- Lambda is used for balancing simplicity while keeping training data fit. ✓

**A7-** Bir lambda değeri seçerken amaç, basitlik ve eğitim verisi (training data) uyumu arasında doğru dengeyi sağlamaktır:

- Lambda değerinizi çok yüksekse, modeliniz basit olacaktır, ancak verilerinizde underfitting riskiyle karşı karşıya kalırsınız. Modeliniz, yararlı tahminler yapmak için eğitim verileri hakkında yeterince bilgi edinmeyecektir.
- Lambda değerinizi çok düşükse, modeliniz daha karmaşık olacaktır ve verilerinizde overfitting riskiyle karşı karşıya kalırsınız. Modeliniz, eğitim verilerinin özellikleri hakkında çok fazla şey öğrenecek ve gelecek yeni verilerde genelleme yapamayacaktır.

**Q8- Why and when do we use sparse representation?**

- When data size is large and most of feature value that we are interested in is zero. ✓
- When data size is large and most of feature value that we are interested in is non-zero.
- When data size is small and a few of feature value that we are interested in is zero.
- When data size is small and a few of feature value that we are interested in is non-zero.

**A8-** Varsayalım ki; elinizdeki veri kümesinde, sokak\_adi için değer olarak eklemek istediğiniz 1.000.000 farklı sokak adı var. Yalnızca 1 veya 2 ögenin doğru olduğu 1.000.000 öğeden oluşan ikili bir vektör oluşturmak, bu vektörleri işlerken hem depolama hem de hesaplama süresi açısından çok verimsiz bir gösterimdir. Böyle durumlardaki ortak bir yaklaşım, yalnızca sıfır olmayan değerlerin depolandığı seyrek gösterim (sparse representation) kullanmaktır.

**Q9- Why do we use the sigmoid function?**

- To activate the neurons separately.
- To map the negative inputs as strongly negative.
- To introduce a new parameter as a slope of the negative part of the function.
- To predict the probability between the range of 0 and 1. ✓

**A9-** Sigmoid aktivasyon fonksiyonu, [0-1] aralığında çıktısı olan olasılık tahminlerinde kullanılır. Fonsiyonun negative kısmının eğimi olarak yeni bir parameter eklemek için Parameterised ReLU aktivasyon fonksiyonu, negative girdileri oldukça güçlü negatifler olarak belirlemek için tanh aktivasyon fonksiyonu ve nöronları tek tek active etmek için de ReLU aktivasyon fonksiyonu kullanılır.

**Q10- Which of the following statements about choosing a good feature is false?**

- Should be avoided rarely used discrete feature values
- Should be preferred clear and obvious meanings
- The definition of a feature should be changed over time ✓
- Should not be mixed "magic" values with actual data

**A10-** İyi bir özellik seçiminde, tanımlanan bir özelliğin adı zamanla değiştirilmemelidir. Bu durumda algoritma, her seferinde bu özelliği yeni bir özellik gibi algılayıp ona göre değerlendirmeler yapacak ve çıkan sonuç her seferinde istenenden daha uzak olacaktır.