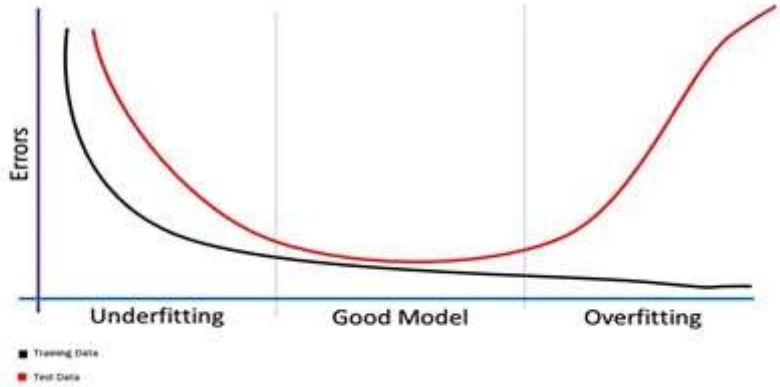
Makine Öğrenmesinde Overfitting / Underfitting / Best Fitting Kavramları

Berna Taş · Follow
Published in Kodluyoruz
5 min read · Sep 9, 2020



Herkese merhabalar 🙂

de aynı durumun söz konusu olacağını düşünürler. Fakat modeli çalıştırdıklarında aslında işlerin pek de doğru gitmediğinin farkına varırlar. Çünkü yeni veriler tahmin edilirken model eğitimden çok daha kötü performans gösterir. Bu çok önemli bir sorundur. Nedeni genellikle bu modellerin eğitim verilerine fazla takılmış (overfitting) olmasıdır. Tam tersi bir durumda söz konusu olabilir(underfitting), şimdi bu konulardan bahsedeceğim.

Buradaki asıl nedeni şöyle de ifade edebiliriz: **Modelin bilgiyi doğru bir şekilde genelleştirememesi**. Bir önceki yazımda genelleştirmeden bahsetmiştim. Buradan genellemeden (generalization)neyi kastettiğime bakabilirsiniz. İyi bir makine öğrenimi modelinin amacı, verileri iyi bir şekilde genelleştirmektir. Bu, gelecekte modelin hiç görmediği veriler hakkında tahminlerde bulunmamızı sağlar.

1. Overfitting

Makine öğrenimi modellerinin amacının yeni verileri doğru bir şekilde tahmin etmek veya çıkarmak olduğu unutulmamalıdır. Başka bir deyişle, yeni verilere yansıtılabilecek genel kalıpları eğitmek ve elde etmektir.

Bu kavram makine öğrenimindeki anahtar kavramlardan biridir. Aşırı uydurma (overfitting), modelimiz eğitim verisini o kadar iyi ezberliyor ki test verisine iyi bir genelleme yapamıyor. Aşırı takma, eğitim veri setinin "aşırı" öğrenildiğini gösterir ve yeni girdi verilerini anlama imkânsızlığı gösterir.

1.1 Aşırı takılmayı (Overfitting) nasıl önleyebiliriz?

Overfit durumuna düşmemek için alınabilecek önlemleri şu şekilde sıralayabilirim:

→ *Cross Validation:* Görünmeyen verilerdeki model doğruluğunu tahmin etmek için uygulanan Machine Learning'de önemli bir tekniktir. Veri setini k tane parçaya ayırarak eğitimi yapar, bu k parçadan 1 parçayı test için kullanır, bu parça her seferinde bir önceki iterasyondan farklı olur, bu yüzden modelimiz sürekli yeni test seti ile test edilmiş olur.

→ *Daha fazla veri ile eğitim (Training with more data)*: Örnek sayımızı artırmak verimizdeki target ile feature arasında ki ilişkiyi daha rahat anlamamızı sağlamayabilmektedir. Daha fazla veriye sahip olan algoritmanın, daha fazla veri türünü dikkate alarak daha iyi genelleme olasılığı daha yüksektir.

→ *Removing Features*: Feature setimizden alakasız feature'ları çıkartarak target-feature ilişkisini daha net bir hale getirebilmekteyiz.

→ *Regularization*: Düzenleme, modelin daha iyi genelleşmesi için öğrenme algoritmasında küçük değişiklikler yapan bir tekniktir. Bu da modelin görünmeyen verilerdeki performansını artırır. Genellikle L1 ve L2 olmak üzere iki türü bulunmaktadır.

→ *Ensembling*: Birbirinden ayrı modelleri bir arada kullanmamıza olanak sağlayan ML metodudur. Böylece modelimiz daha karmaşık yapılı örnekler ile overfit olmadan çalışabilir.

→ *Early stopping:* Çok fazla dönem, eğitim veri kümesinin aşırı sığmasına neden olabilirken, çok azı bir yetersizlik modeliyle sonuçlanabilir. Bir Makine Öğrenimi modelini yinelemeli olarak eğittiğinizde, birkaç yinelemeye kadar modelin performansının arttığını fark edeceksiniz. Belirli bir noktadan sonra, yineleme sayısını artırırsanız, model eğitim veri kümesinde daha iyi performans gösterir, ancak model aşırı yüklenir ve test veri kümelerinde düşük performans gösterir. Bu nedenle, modele fazla uymadan önce modelinizin eğitim tekrarlarını durdurmalısınız.

2. Underfitting

Train error yüksekse, makine öğrenmesi modelimizde öğrenme güçlüğü vardır. Buna İngilizcede *undefitting* denir. Underfitting, eğitim veri kümesini modelleyemeyen veya yeni veri kümesini genelleştiremeyen bir modeli ifade eder. Bu uygun bir model yapısı değildir.

2.1 Modelimizin underfitting olmasının nedenleri

→ *Basit model yapısı*: Oluşturmuş olduğunuz model o kadar basittir ki, girdi ve çıktı verileri arasındaki ilişkiyi tam olarak öğrenemez. Yüksek eğitim hatası (high train error), makine öğrenme modelinin çok basit olduğunu gösterir.

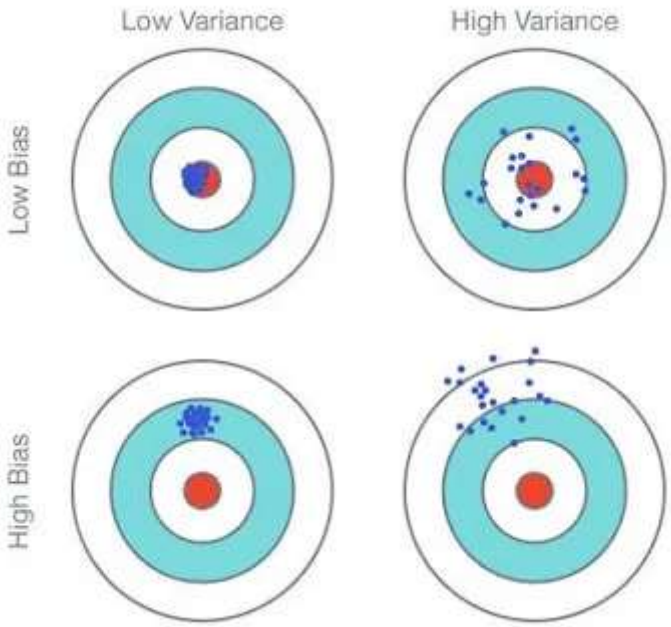
→ *Eksik veri*: Underfitting için bir başka neden veri eksikliği olabilir. Örneğin, sonuçlar birden fazla değişkene bağlıyken siz eğitimde sadece bir değişken üzerinden öğrenmeye çalışırsanız eğitim performansınız düşük olacaktır.

→ Yetersiz veri: Modellerimizin yüksek başarı oranları vermesini bekliyorsak veri miktarının yeterli olması gerekmektedir. Elimizde yeteri kadar veri yoksa modelimiz giriş verileri ve sonuçlar arasında bir örüntü bulmakta zorlanacaktır.

→ Gürültülü veri: Veriler çok gürültülü ise, makine öğrenme teknikleri öğrenme güçlüğü çekecek ve genelleme yapamayacaktır. Modelimizin elimizdeki verilerin rastgele oluşturulduğunu düşünmemesi için gürültülü verileri azaltmaya çalışmalıyız.

3. Bias — Variance

Çok basit bir şekilde yüksek bir yanlılık (bias) modelin underfit olduğunu, yüksek bir varyans (variance) modelin overfit olduğunu göstermektedir.

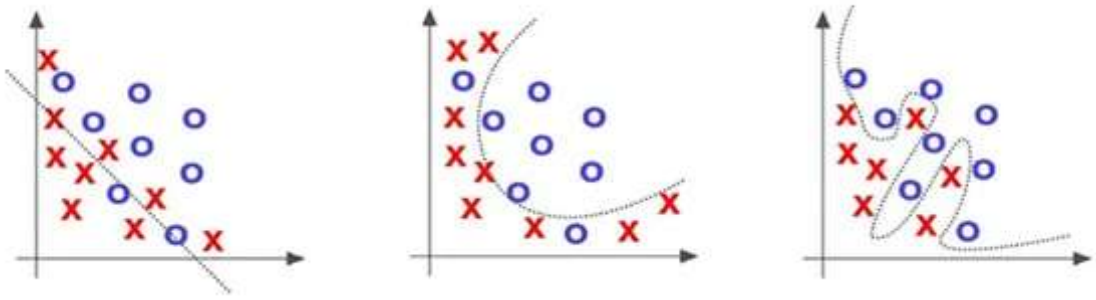


Source: https://towardsdatascience.com/regularization-the-path-to-bias-variance-trade-off-b7a7088b4577

- Yüksek yanlılık (bias) durumu: Modelimiz underfitting olayından etkilenmektedir. Bir çözüm olarak, analiz edilecek özelliklerin sayısını veya algoritmanın karmaşıklığını arttırmayı deneyebiliriz.
- Yüksek varyans (variance) durumu: Modelimiz overfitting durumundan etkilenmektedir. Modelimiz train verilerine çok iyi uyuyor ancak test verilerini doğru bir şekilde tahmin edemiyor. Bu nedenle test hatası train hatasından önemli ölçüde daha yüksek gelmektedir.

4. Nasıl değerlendirilir?

Bir modelin overfitting ve underfitting durumunu değerlendirmek için çeşitli yöntemler vardır. Yöntemlerden biri, eğitim ve test hata grafiklerine bakmaktır. Model hata grafikleri, eğitmek (train) için kullanılan verilerde ve modeli doğrulamak için kullanılan verilerde (test) temsil edilebilir. İdeal olarak, her iki hata da olabildiğince yakın olmalıdır.



Sol taraftaki grafiğe baktığımızda, çizginin grafikte gösterilen tüm noktaları kapsamadığını görüyoruz. Böyle bir modelde veriler underfitting olma eğilimindedir, buna yüksek yanlılık (bias) denir.

Sağdaki grafikte gösterilen çizginin tüm noktaları kapsadığını söyleyebiliriz. Böyle bir durumda yani tüm noktaları kapsadığı için iyi bir grafik olduğunu düşünebilirsiniz, ancak bu doğru değil, grafikteki çizgi de gürültü ve aykırı olan tüm noktaları kapsar. Bu model, karmaşıklığı nedeniyle overfitting durumuna sebebiyet vermekte buna yüksek varyans (variance) da denir.

Son olarak ortadaki grafiğe baktığımızda, iyi bir tahmin çizgisi olduğunu söyleyebiliriz. Çünkü grafikteki noktaların çoğunu kapsıyor ve aynı zamanda önyargı veya yanlılık ve varyans arasındaki dengeyi koruyor.

5. Sonuç

Yazım boyunca anlatmış olduğum kavramları kısaca toparlayacak olursam:

Modelimiz eğitim veri kümesinde test veri kümesine göre çok daha iyi sonuç veriyorsa, büyük olasılıkla burada overfitting durumunun söz konusu olduğunu söyleyebiliriz. Örneğin, modelimizin eğitim başarısı % 90'larda seyrederken test setinde % 50 ve altında bir doğruluk gerçekleşmiş olabilir. Yani görünmeyen veri kümesinde modelimiz iyi performans göstermedi. Modelimiz test veri kümesinde eğitim veri kümesinden çok daha iyi sonuç veriyorsa, muhtemelen underfitting durumu söz konusu olmaktadır. Modelimiz hem eğitim hem de test veri kümelerinde iyi sonuç veriyorsa (best fitting) en uygun modeli elde ettiğimizi söyleyebiliriz. Örneğin, modelimiz eğitim veri kümesinde % 90'larda başarı gösterirken, test veri kümesinde % 80 -% 95 doğruluk gerçekleştirmiştir.

Makine Öğrenimi modelleri geliştirirken, modellerimizin iyi sonuçlar vermesini istiyorsak bu tanımların her birinin bilinmesinde oldukça fayda vardır. Umarım bu

makale, yeni verilere iyi genelleşen modeller yapmanın önemini anlamanız açısından faydalı olmuştur. Hep birlikte doğru ve yararlı kaynaklar oluşturabilmek adına hatalı veya eksik bulduğunuz bir yer olursa yorum olarak belirtebilirsiniz. Bir sonraki yazımda görüşmek üzere. Sağlıklı günler dilerim 🙂

Machine Learning | Overfitting | Underfitting | Bias | Yazılımcı

Follow

Written by Berna Taş

396 Followers · Writer for Kodluyoruz

Business Intelligence Consultant https://www.linkedin.com/in/bernaatas/

More from Berna Taş and Kodluyoruz

Berna Taş

Doğrusal Regresyon (Linear Regression)

Herkese merhabalar, bu yazımda Doğrusal Regresyondan bahsedeceğim. Umarım faydalı bir içerik olur. Keyifli okumalar dilerim :)

✨ · 4 min read · May 15, 2020

288 3

kodluyoruz in Kodluyoruz

Genç Yazılımcılar için CV ve Ön Yazı Hazırlama Tavsiyeleri

CV ve ön yazı işe giriş süreçlerinde fark yaratıyor. Özellikle sektöre yeni adım atan Genç yazılımcılar için cv ve ön yazı kariyerlerinin...

5 min read · Jun 21, 2019

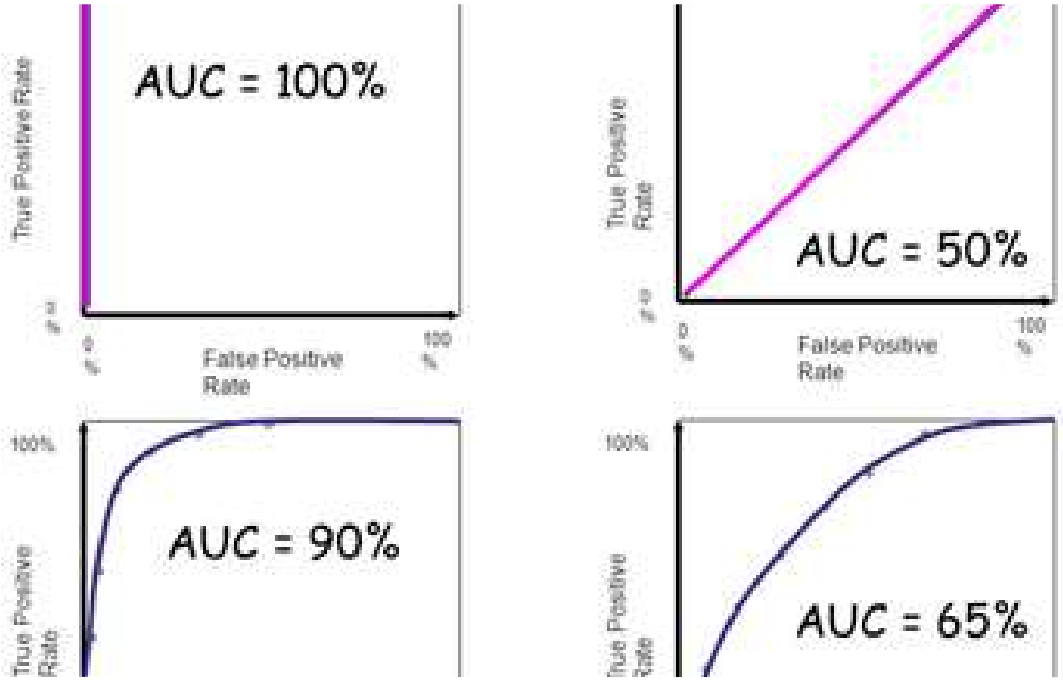
344

Semih Elitaş in Kodluyoruz

ASP.NET Core Identity ile Rol Bazlı Üyelik Sistemi Oluşturmak & Kullanıcıyı Özelleştirmek

Herkese selamlar! Umarım hepiniz iyi ve sağlıklısınızdır. İçinde bulunduğumuz virüs sürecince boş durmamak adına, yeni başlayan arkadaşlar...

9 min read · Jul 19, 2020



Berna Taş

Roc Eğrisi ve Eğri Altında Kalan Alan (Auc)

Bu yazımda oluşturduğumuz sınıflandırma modellerinin performansını nasıl değerlendirebileceğimizi ROC ve AUC eğrilerinden bahsederek...

5 min read · Nov 26, 2019

319 1

See all from Berna Taş

See all from Kodluyoruz

Recommended from Medium

Automata in **AI monks.io**

Improving Stock Price Forecasting by Feature Engineering

In this article, I want to share with you how I tackled the problem of predicting the value of the stock at the next day's close, using...

10 min read · Jul 18

93 2

Alexandru Lazar in ILLUMINATION

Ten Habits that will get you ahead of 99% of People

Improve your life and get ahead of your peers in 10 simple steps

9 min read · Nov 18

12.7K 224

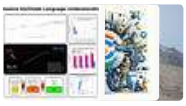
Lists

Predictive Modeling w/ Python

20 stories · 711 saves

Practical Guides to Machine Learning

10 stories · 814 saves

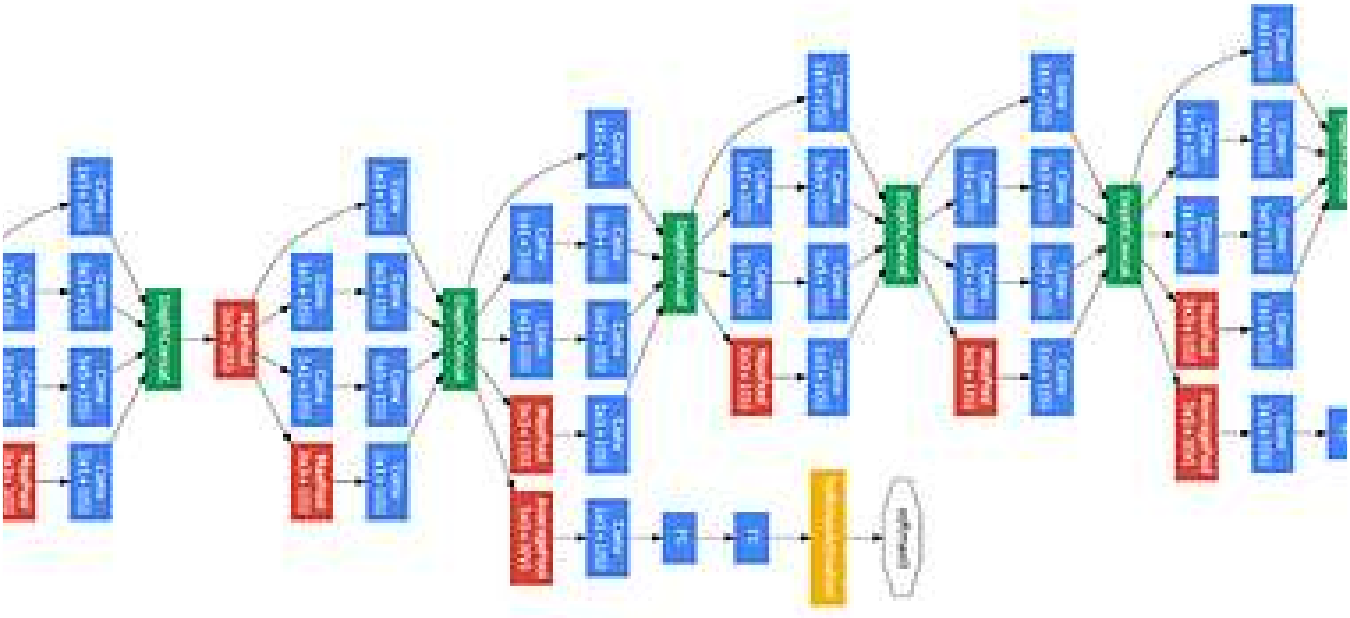


Natural Language Processing

1013 stories · 495 saves

The New Chatbots: ChatGPT, Bard, and Beyond

12 stories · 245 saves

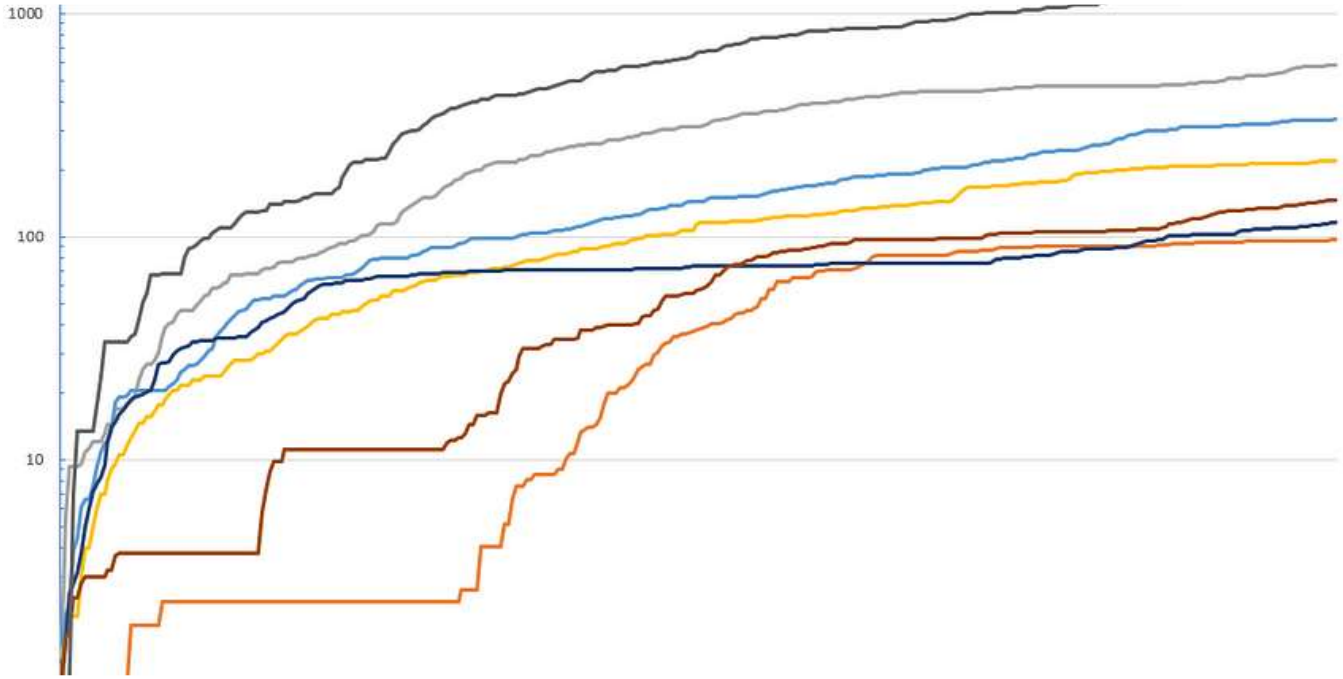


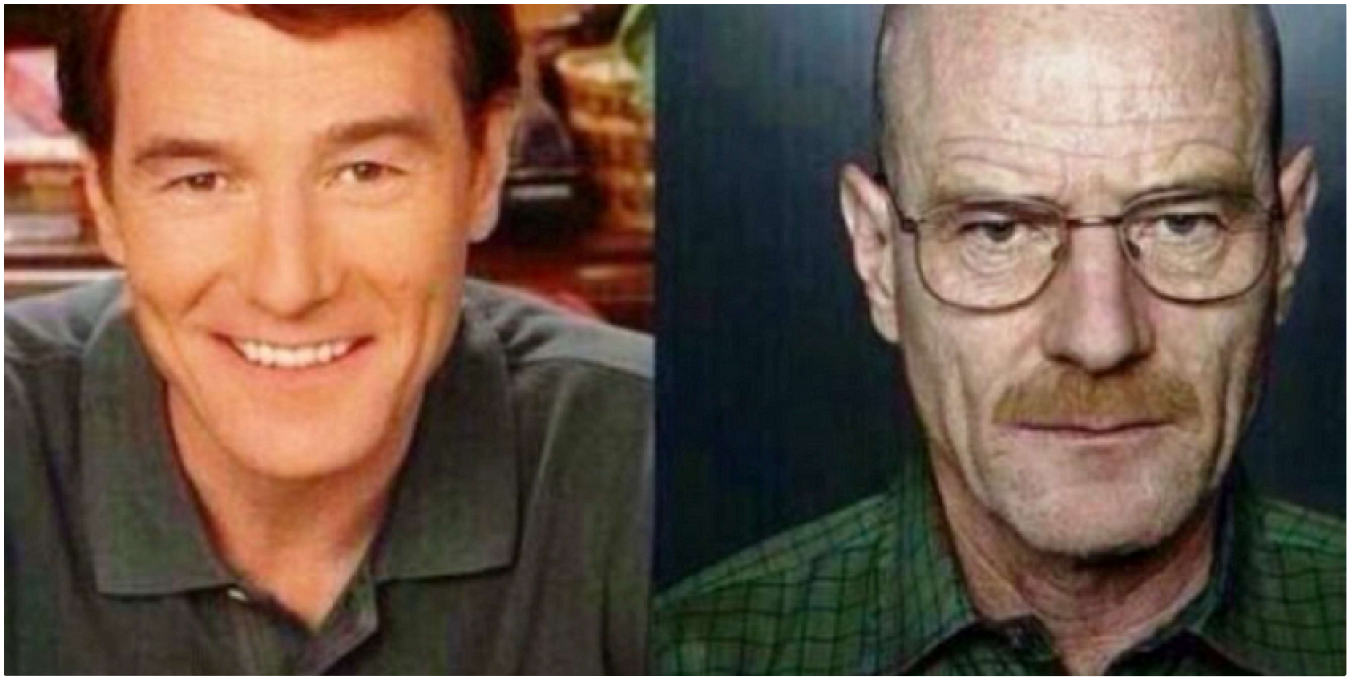
Everton Gomede, PhD

Exploring GoogLeNet: A Revolutionary Deep Learning Architecture

Introduction

7 min read · Oct 2





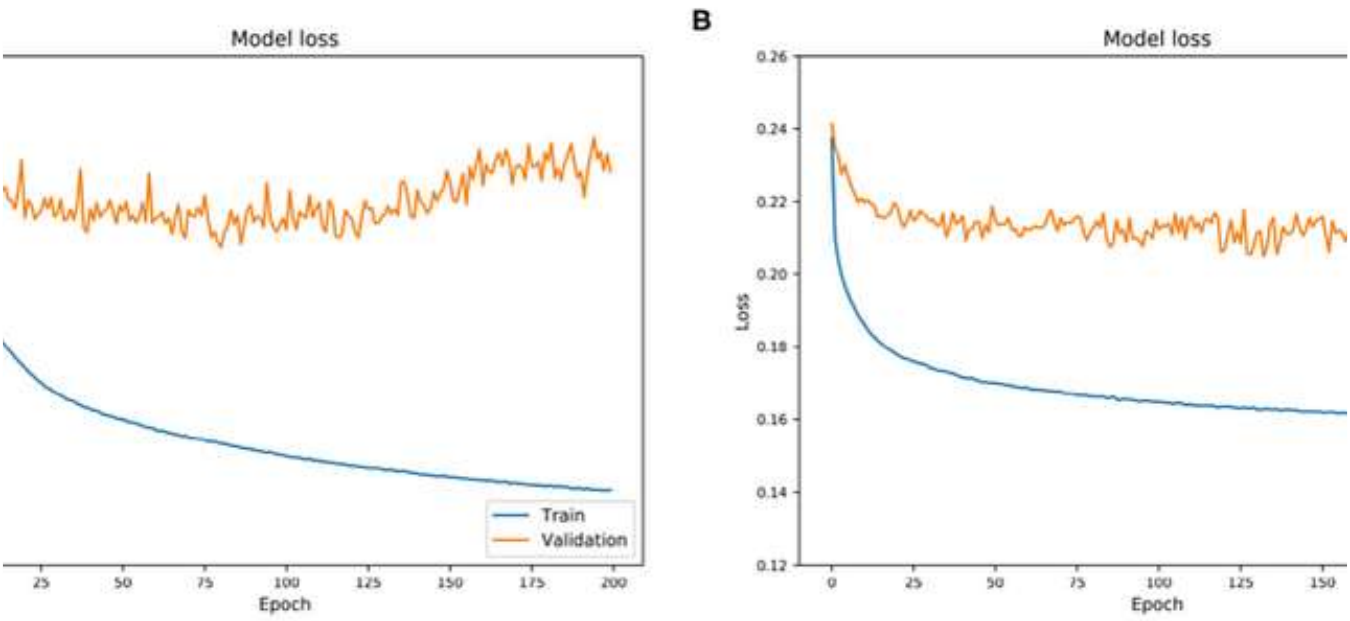
David Goudet

This is Why I Didn't Accept You as a Senior Software Engineer

An Alarming Trend in The Software Industry

5 min read · Jul 26

6.8K 72



Koushik

What is Overfitting and Underfitting , and how to deal with it step by step?

In machine learning, it is common to face a situation when the accuracy of models on the validation data would peak after training for a...

5 min read · Sep 20

 68