Sıra desen madenciliği ve Rassal orman modeli

Sıra Desen Madenciliği

- Sıra desen madenciliği, belirli bir sırayla gelen istatiksel olarak veri örnekleri arasındaki ilgili örüntüleri bulmaya çalışır. Müşterilerin teknoloji market alışverişi verilerine göre son 3 ayda sırasıyla önce bilgisayar sonra CD-ROM son olarak dijital kamera satın almaları, tıbbi tedaviler, doğal felaketler(deprem), DNA dizilişi ve gen yapısı sıralı örüntü madenciliği ile ilgilidir. Mesela Internet şubesinde yapılan işlemler bir sıraya göre yapıldığı için sıralı örüntü madenciliği içerisinde yer almaktadır. sıralı örüntü madenciliğinde arka arkaya yapılan işlemler göz önüne alınır.

Rastgele Orman Sınıflandırıcısı:

- Rastgele orman, adından da anlaşılacağı gibi, bir topluluk olarak çalışan çok sayıda bireysel karar ağacından oluşur. Rastgele ormandaki her bir ağaç bir sınıf tahmini verir ve en çok oyu alan sınıf, modelimizin öngörüsü haline gelir (aşağıdaki şekle bakın).

Rastgele orman

- Rastgele ormanın ardındaki temel kavram basit ama güçlü bir kavramdır - kalabalıkların bilgeliği. Veri biliminde konuşursak, rastgele orman modelinin bu kadar iyi çalışmasının nedeni şudur: Bir komite olarak faaliyet gösteren çok sayıda görece ilişkisiz model (ağaç), münferit kurucu modellerin herhangi birinden daha iyi performans gösterecektir. Modeller arasındaki düşük korelasyon anahtardır. Tıpkı düşük korelasyonlu yatırımların (hisse senetleri ve tahviller gibi) bir araya gelerek parçalarının toplamından daha büyük bir portföy oluşturması gibi, ilişkisiz modeller, bireysel tahminlerin herhangi birinden daha doğru olan topluluk tahminleri üretebilir.

Neden rassal orman?

- Bu harika etkinin nedeni, ağaçların birbirlerini kendi hatalarından korumalarıdır (sürekli aynı yönde hata yapmadıkları sürece). Bazı ağaçlar yanlış olabilirken, diğer birçok ağaç haklı olacaktır, bu nedenle bir grup olarak ağaçlar doğru yönde hareket edebilecektir. Bu nedenle, rastgele ormanın iyi performans göstermesi için ön koşullar şunlardır: Özelliklerimizde bazı gerçek sinyaller olması gerekir, böylece bu özellikler kullanılarak oluşturulan modeller rastgele tahmin etmekten daha iyi sonuç verir. Tek tek ağaçların yaptığı tahminlerin (ve dolayısıyla hataların) birbirleriyle düşük korelasyonlara sahip olması gerekir.

Örnek

- İlişkisiz sonuçların neden bu kadar büyük olduğuna dair bir örnek: İlişkisiz birçok modele sahip olmanın harika etkileri o kadar kritik bir kavramdır ki, şu oyunu oynadığımızı hayal edin: Bir sayı üretmek için tekdüze dağıtılmış rasgele sayı üreteci kullanıyorum. Oluşturduğum sayı 40'tan büyük veya ona eşitse, kazanırsınız (yani% 60 zafer şansınız olur) ve size biraz para öderim. 40'ın altındaysa ben kazanırım ve siz bana aynı miktarı ödersin. Şimdi size aşağıdaki seçenekleri sunuyorum. Şunlardan birini yapabiliriz: Oyun 1 - 100 kez oynayın, her seferinde 1 $ bahis yapın. Oyun 2 - 10 kez oynayın, her seferinde 10 $ bahis yapın. Oyun 3 - bir kez oynayarak 100 $ bahis yapın.

Örnek

- Hangisini seçerdin? Her oyunun beklenen değeri aynıdır: Beklenen Değer Oyunu 1 = (0.60 * 1 + 0.40 * -1) * 100 = 20 Beklenen Değer Oyunu 2 = (0.60 * 10 + 0.40 * -10) * 10 = 20 Beklenen Değer Oyunu 3 = 0.60 * 100 + 0.40 * -100 = 20

Örnek

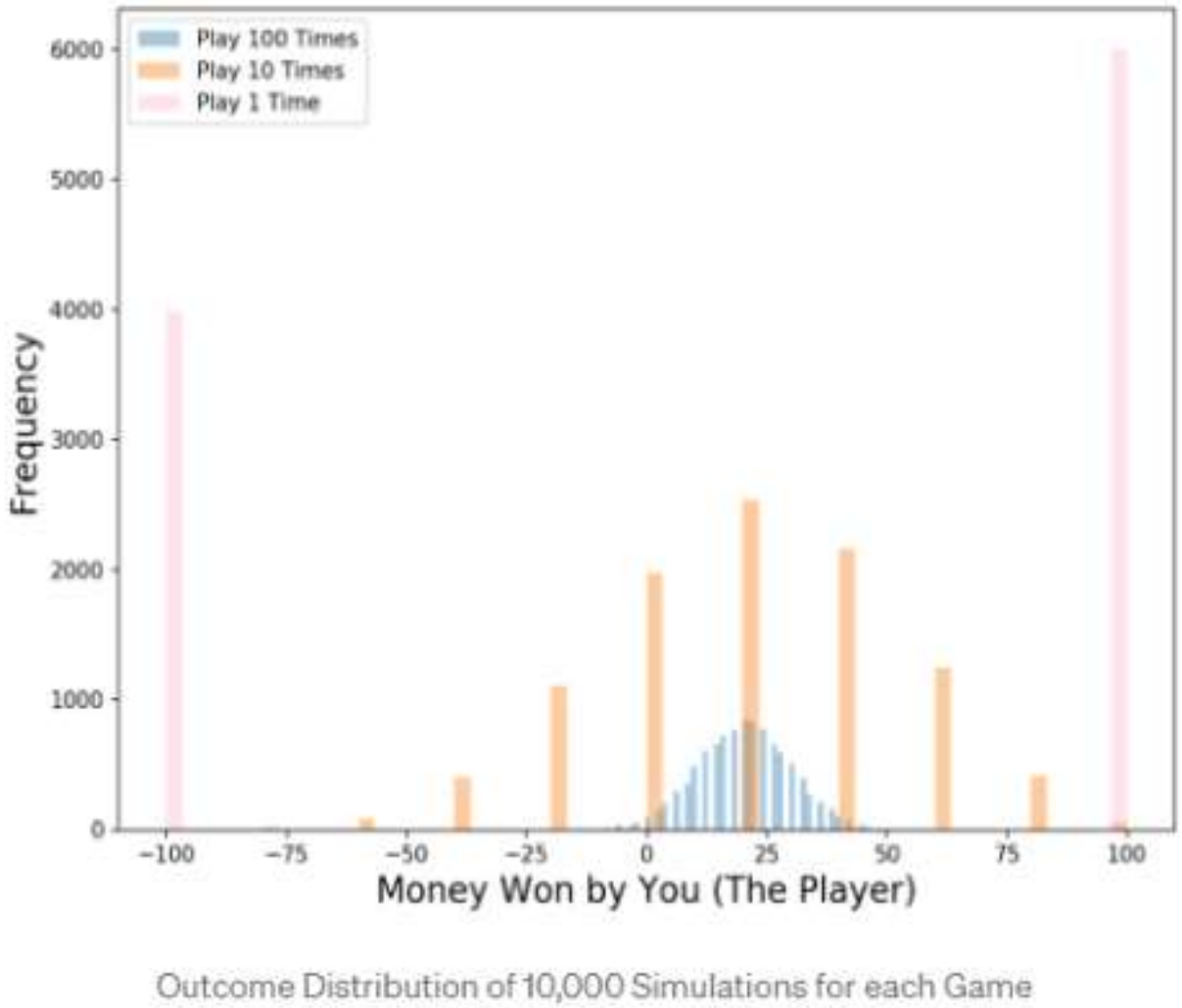
- Dağılımlar ne olacak? Sonuçları bir Monte Carlo simülasyonu ile görselleştirelim (her oyun türü için 10.000 simülasyon çalıştıracağız; örneğin, 1. Oyundaki 100 oyunun 10.000 katını simüle edeceğiz). Soldaki tabloya bir göz atın - şimdi hangi oyunu seçerdiniz? Beklenen değerler aynı olsa bile, sonuç dağılımları, pozitif ve dardan (mavi) ikiliye (pembe) doğru büyük ölçüde farklıdır.

Örnek

- Oyun 1 (100 kez oynadığımız yer), biraz para kazanmak için en iyi şansı sunuyor - yürüttüğüm 10.000 simülasyondan% 97'sinde para kazanıyorsunuz! Oyun 2'de (10 kez oynadığımız yerde) simülasyonların% 63'ünde para kazanırsınız, ciddi bir düşüş (ve para kaybetme olasılığınızda ciddi bir artış). Ve sadece bir kez oynadığımız 3. Oyun, beklendiği gibi simülasyonların% 60'ında para kazanıyorsunuz.

Örnek

- Dolayısıyla, oyunlar aynı beklenen değeri paylaşsa da, sonuç dağılımları tamamen farklıdır. 100 $ 'lık bahsimizi farklı oyunlara ne kadar çok bölersek, para kazanacağımıza o kadar güvenebiliriz. Daha önce de belirtildiği gibi, bu işe yarar çünkü her oyun diğerlerinden bağımsızdır. Rastgele orman aynıdır - her ağaç, önceki oyunumuzdaki bir oyun gibidir. Daha fazla oynadığımızda para kazanma şansımızın nasıl arttığını gördük. Benzer şekilde, rastgele bir orman modeliyle, modelimizdeki ilişkisiz ağaçların sayısı ile doğru tahminler yapma şansımız artar.



Outcome Distribution of 10,000 Simulations for each Game

Modellerin birbirini çeşitlendirmesini sağlamak:

- Öyleyse rastgele orman, her bir ağacın davranışının modeldeki diğer ağaçlardan herhangi birinin davranışıyla çok fazla ilişkili olmamasını nasıl sağlar? Aşağıdaki iki yöntemi kullanır: Torbalama (Bootstrap Aggregation) - Karar ağaçları, eğitildikleri verilere karşı çok hassastır - eğitim setinde yapılan küçük değişiklikler, önemli ölçüde farklı ağaç yapılarına neden olabilir. Rastgele orman, her bir ağacın veri kümesinden değiştirilerek rasgele örneklemesine izin vererek bundan yararlanır ve farklı ağaçlarla sonuçlanır.

Modellerin birbirini çeşitlendirmesini sağlamak:

- Bu işlem torbalama olarak bilinir. Torbalama ile eğitim verilerini daha küçük parçalara ayırmadığımıza ve her ağacı farklı bir yığın üzerinde eğitmediğimize dikkat edin. Bunun yerine, N büyüklüğünde bir örneğimiz varsa, yine de her ağaca N boyutunda bir eğitim seti besliyoruz (aksi belirtilmedikçe). Ancak orijinal eğitim verileri yerine, değiştirilmiş N boyutunda rastgele bir örnek alıyoruz. Örneğin, eğitim verilerimiz [1, 2, 3, 4, 5, 6] ise, ağaçlarımızdan birine aşağıdaki listeyi verebiliriz [1, 2, 2, 3, 6, 6]. Her iki listenin de altı uzunluğunda olduğuna ve "2" ile "6" nın ikisinin de ağacımıza verdiğimiz rastgele seçilmiş eğitim verilerinde tekrarlandığına dikkat edin (çünkü değiştirme ile örnekleme yapıyoruz).