Çözümler : 16 Kasım 2015, Şadi Evren ŞEKER

T.C. İstanbul Ticaret Üniversitesi
Veri Madenciliğine Giriş Dersi
Vize İmtihanı

Her soru eşit puandır, sınav açık kitaptır.

1) Veri Ambarı (data warehouse) ve KDD (knowledge discovery in databases, veri tabanlarında bilgi keşfi) ve İş Zekası (Business Intelligence) kavramlarının ilişkilerini açıklayınız.

İş zekası, KDD aşamalarının sonuncusunu/sonucunu teşkil eder ve KDD aşamalarının daha hızlı, verimli ve başarılı çalışması için veri ambarlarına ihtiyaç duyulur.

2) OLAP (Online Analytical Processing – Çevrimiçi Analitik İşleme) , Veri Ambarı (Data Warehouse) ve Büyük Veri (Big Data) kavramları arasındaki ilişkiyi açıklayınız. Büyük veri dünyasında OLTP (Online Transaction Processing-Çevrimiçi Hareket İşleme) veri tabanları (database) için neler söyleyebilirsiniz başlıklar halinde yazınız.

OLAP teknolojisi, OLTP teknolojisinden farklı olarak analiz ağırlıklı işler için tasarlanmıştır ve bu tasarımın uygulama alanları veri ambarlarıdır. Büyük veri ise bütün veri tabanının bir veri ambarı olduğu ve aynı zamanda OLAP yaklaşımına göre tasarlandığı ortamdır.

3) Aşağıda verilen veri kümesi için hangi sınıflandırma algoritmasını tavsiye ederdiniz, sebepleri ile açıklayınız (sınıflandırma işlemi yapmayınız sadece hangi algoritmayı tercih edeceğinizi ve sebeplerini yazınız). Bildiğiniz diğer sınıflandırma algoritmalarından en az 4 tanesini neden tercih etmediğiniz açıklayınız. Hangi özelliğe (attribute) göre sınıflandırma yaptığınızı belirtin.

Boy	Kilo	Cinsiyet	Yaş	Eğitim
180	80	erkek	30	Üniversite
170	85	erkek	40	Üniversite
180	80	erkek	32	İlkokul
160	40	-	25	İlkokul
165	55	kadın	28	-
205	140	erkek	27	Üniversite
175	58	kadın	20	Lise
185	65	kadın	18	Lise
158	48	kadın	29	y.lisans
190	120	erkek	35	y.lisans

Soruda iki tip sınıflandırma yapılabilir, bunlardan birisi eğitim diğeri ise cinsiyettir. Çünkü diğer özellikler sayısal olup sınıflandırma için doğrudan uygun değildir. Eğitim veya cinsiyete göre sınıflandırma için bir özellik seçilirken daha doğru olanı cinsiyeti seçmektir çünkü eğitim için her sınıftan yeterli örnek olmadığı görülebilir (sınıflarda sadece 2 veya 3 örnek bulunması yetersiz veri dolayısıyla problem oluşturacaktır). C
insiyete göre sınıflandırma için KNN algoritmasının en iyi sonucu vermesi beklenir. Bunun sebebi boy ve kilo ile cinsiyet arasında doğrudan ilişki bulunmasıdır. Eğitim ile de yaş arasında bağlantı bulunması beklenir ancak örnekte veriye bakıldığında

böyle bir birliktelik görülmemektedir. (Not: KNN algoritması ile K=3 için cinsiyet üzerinden sınıflandırma yapılması %100 başarı vermektedir.)
Diğer algoritmaların seçilmeme sebebi: OneR ve ZeroR algoritmaları çok özellikli (birden fazla kolon içeren) durumlarda başarısız çalışması beklenen algoritmalardır. Naive Bayes algoritması bu örnekte kullanılamaz çünkü yapabileceği sınıflandırma için doğrudan kullanabileceği tek özellik eğitimdir. Yani cinsiyet ve eğitim arasında olasılıksal bir hesaplama yapabilecektir (diğer özellikler için verinin quantize edilmesi gerekir ve doğrudan kullanılamaz). Dolayısıyla naive bayes uygun sınıflandırma algoritması değildir.
Karar ağaçları eksik veri ile çalışmak için uygun değildir. Imputation yapılarak eksik veri silinirse başarılı çalışabilir (örneğin 65 kilo ve altı kadın üzeri ise erkek şeklinde)

4) Bir tahmin (prediction) problemini nasıl bir sınıflandırma problemine dönüştürürdünüz açıklayınız. Örnek olarak bir önceki soruda verilen kişilerin yaşlarını doğru tahmini için bildiğiniz sınıflandırma algoritmalarından birisini kullanın ve en alt satırda verilen kişinin yaşını tahmin edin.

Dönüşüm için özellik üzerinde quantization yapılabilir. Yaşlar belirli yaş aralıklarında gruplanabilir. Örneğin genç, ortayaş ve yaşlı gibi sınıflara bölünerek hangi sınıfta olduğunun tahmin edilmesi aslında yaş aralığının tahmini olarak düşünülebilir. Örneğin çözümü için 5 yaşlık quadrantlar alınırsa tablo aşağıdaki şekilde dönüşecektir (15'ten 40'a kadar olan her 5 yaşlık grup için bir sayı atanmıştır):

Boy	Kilo	Cinsiyet	Yaş	Eğitim
180	80	erkek	3	Üniversite
170	85	erkek	5	Üniversite
180	80	erkek	4	İlkokul
160	40	-	2	İlkokul
165	55	kadın	3	-
205	140	erkek	3	Üniversite
175	58	kadın	1	Lise
185	65	kadın	1	Lise
158	48	kadın	3	y.lisans
190	120	erkek	4	y.lisans

Sorunun yeni haliyle bir sınıflandırma problemi olduğu söylenebilir. Yukarıdaki tablodan kural çıkarımı yapılırsa (rule based classification algoritması) aşağıdaki kurallar yazılarak sınıflandırma yapılmış olunur (sınavda farklı bir algoritma seçilebilirdi):

- lise -> 1
- kadın ve y.lisans veya erkek ve üniversite -> 3
- erkek ve üniversite -> 5
- ilkokul ve 80 kilo'dan küçük -> 2
- erkek ve (ilkokul veya y.lisans) -> 4

Çözümler : 16 Kasım 2015, Şadi Evren ŞEKER

5) 3. Soruda verilen veri kümesi için daha başarılı sınıflandırma için hangi ön işleme (preprocessing) aşamalarını tavsiye ederdiniz?

Eksik verilerin tamamlanması, bazı sınıflandırma algoritmaları için sayısal verilerin quantization'ı (numeric -> nominal dönüşümü).