Eskişehir Osmangazi Üniversitesi
Fen Bilimleri Enstitüsü
İnşaat Mühendisliği Anabilim Dalı

KESİKLİ TERCİH MODELLERİ

Discrete Choice Models

Dr. Kadir Berkhan AKALIN

3

İSTATİSTİKSEL ANALİZ YÖNTEMLERİ

Toplanan verilerin anlamlandırılması ve açıklanması amacıyla kullanılan yöntemlerdir.

<table>
<tr><th colspan="2">BETİMSEL İSTATİSTİK</th><th colspan="2">ÇIKARIMSAL İSTATİSTİK</th></tr>
<tr><th>KAPSAM</th><th>AMAÇ</th><th>KAPSAM</th><th>AMAÇ</th></tr>
<tr><td>VERİ DÜZENLEME</td><td rowspan="3">VERİNİN AÇIKLANMASI</td><td>TAHMİN</td><td rowspan="3">ANAKÜTLE KARAKTERİSTİKLERİ HAKKINDA KARAR VERME</td></tr>
<tr><td>VERİ TANITIMI</td><td rowspan="2">HİPOTEZ TESTİ</td></tr>
<tr><td>VERİ KARAKTERİZASYONU</td></tr>
</table>

OLASILIK (PROBABILTY)

Her olaya 0 ile 1 arasında bir gerçel sayı tahsis eden bir fonksiyondur. Olasılık fonksiyonunun belirtildiği örnek uzaya olasılık uzayı denir. Evrensel sonuç kümesinin (anakütle) sonlu olduğunu, her sonucun meydana gelme olasılığının eşit olduğunu ve birbirini dışladığı (bir zarın atılması gibi) varsayımında A sonucuyla sonuçlanması muhtemel olan n olası sonuç ve m durum kümesi göz önüne alındığında, A sonucunun olma olasılığı aşağıdaki gibi hesaplanır:

$$\boldsymbol{P(A) = \frac{m}{n}}$$

OLASILIK (PROBABILTY)

Örnek 1: 6 yüzlü bir zarın 3 gelme olasılığı

m = {3} (3 gelmesi durumu)
n = {1,2,3,4,5,6} (olası sonuçlar kümesi)

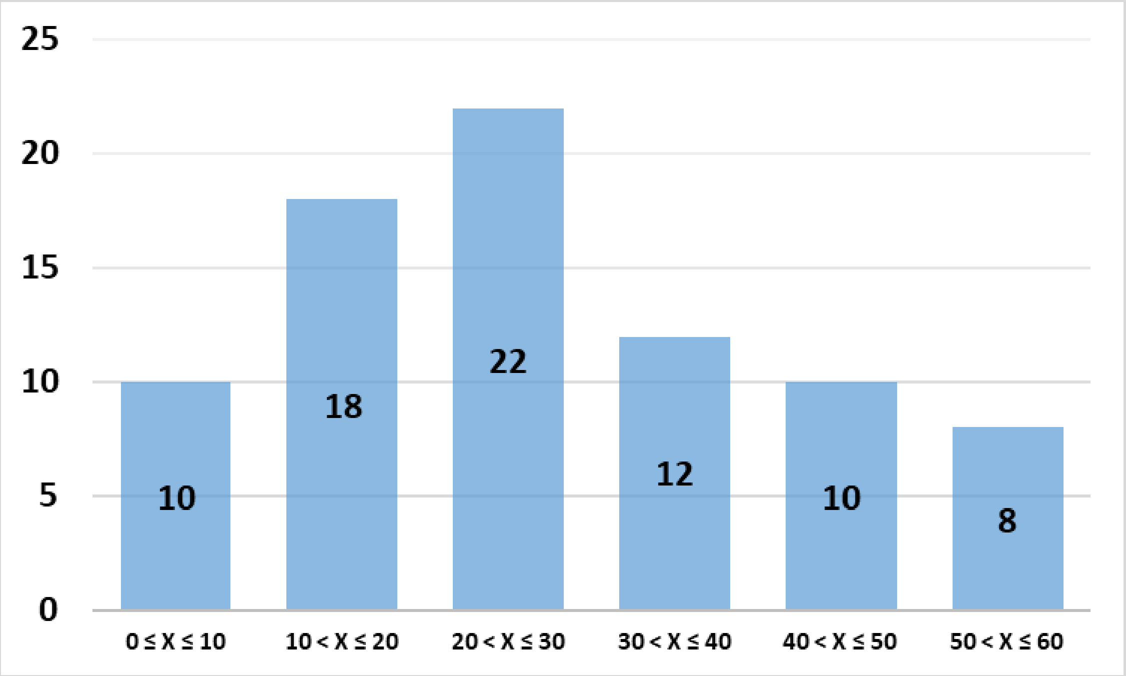
$$P(3) = \frac{1}{6}$$

Örnek 2: 6 yüzlü bir zarın tek sayı gelme olasılığı

m = {1,3,5} (tek sayı gelmesi durumu)
n = {1,2,3,4,5,6} (olası sonuçlar kümesi)

$$P(tek\_sayı) = \frac{3}{6}$$

Yolculuk Süresi (dk)	Frekans	Bağıl Frekans (P)
$0 \leq X \leq 10$	10	10/80 = 0.125
$10 < X \leq 20$	18	18/80 = 0.225
$20 < X \leq 30$	22	22/80 = 0.275
$30 < X \leq 40$	12	12/80 = 0.150
$40 < X \leq 50$	10	10/80 = 0.125
$50 < X \leq 60$	8	8/80 = 0.100
Toplam	**80**	**1.000**



Sonuçlar kümesi sonlu değilse veya herhangi iki sonucun meydana gelme olasılığı eşit değilse, o zaman bağıl frekansları olasılık olarak düşünebiliriz.

RASSAL DEĞİŞKEN (RANDOM VARIABLE)

Değerleri rassal olan ve bu değerler için bir olasılık dağılımı saptama imkânı olan, yani, belirli değerleri belirli olasılıklarla alan değişkendir.

Diğer bir deyişle, bir örnek uzaydaki her rastgele olaya sayısal bir değer atayan bir fonksiyondur. Rasgele değişken gibi büyük harflerle gösterilir. Rastgele değişkenler kesikli rastgele ve sürekli rastgele olmak üzere ikiye ayrılır:

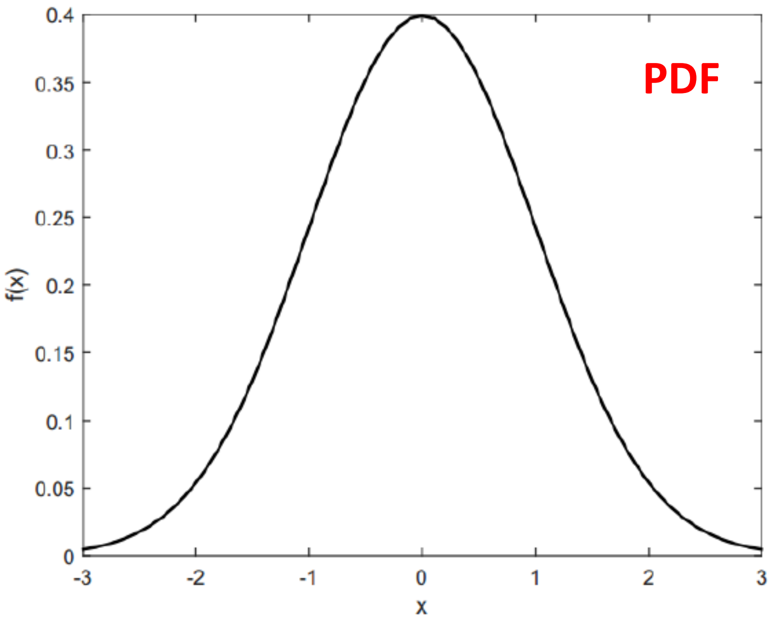
Bir rastgele değişkeninin olası değerlerinin sayısı sonlu veya sayılabilir ise **Sürekli Rassal Değişken,** bir rastgele değişkeninin olası değerleri bir aralıktan yada aralıklar kümesinden oluşuyor ise **Kesikli Rassal Değişken** olarak ifade edilir.

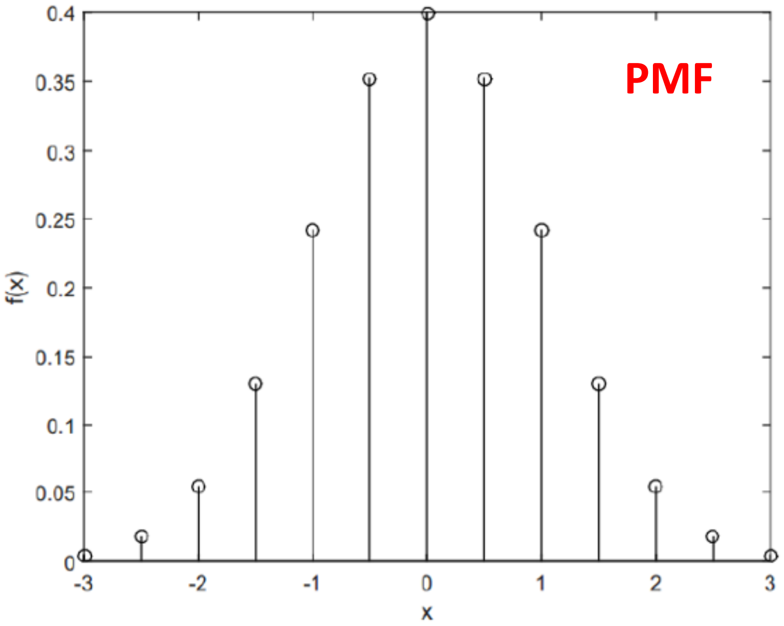
OLASILIK DAĞILIMI (PROBABILITY DISTRIBUTION)

Bir deney için farklı olası sonuçların ortaya çıkma olasılıklarını veren matematiksel fonksiyondur. Bir olasılık dağılımı bir rassal olayın ortaya çıkabilmesi için değerleri ve olasılıkları tanımlar. Değerler olay için mümkün olan tüm sonuçları kapsamalıdır ve olasılıkların toplamı bire eşit olmalıdır.

<u>OLASILIK KÜTLE FONKSİYONU VE OLASILIK YOĞUNLUK FONKSİYONU</u>

Olasılık kütle fonksiyonu (probability mass function –PMF) ve olasılık yoğunluk fonksiyonu (probability density function –PDF), sırasıyla kesikli ve sürekli olasılık dağılımlarını tanımlamak için kullanılır. Bir gözlemin tam olarak bir hedef değere eşit (ayrık) veya hedef değerin etrafında bir aralık içinde (sürekli) olma olasılığının belirlenmesini sağlar.





OLASILIK KÜTLE FONKSİYONU (PMF)

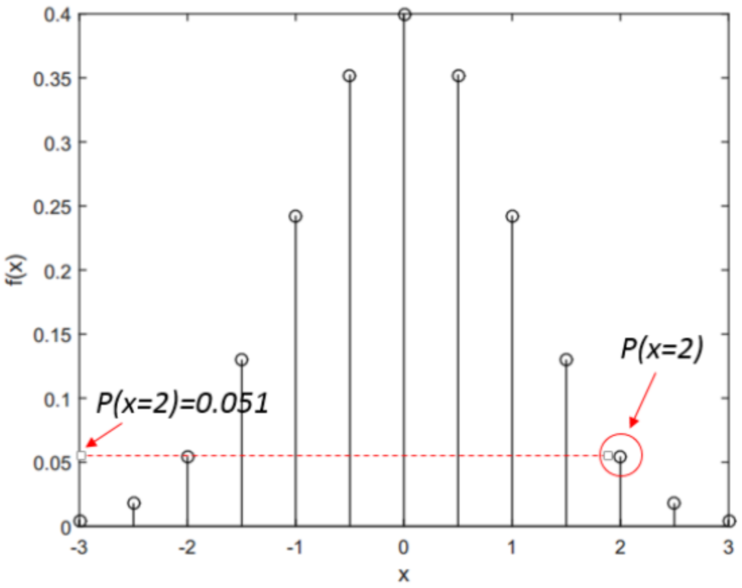
X kesikli bir rastgele değişken olduğunda;

$f_X(x) = P(X = x)$ fonksiyonu rastgele değişkenin olasılık fonksiyonu veya olasılık kütle fonksiyonudur.

Olasılık fonksiyonu aşağıdaki koşulları sağlamalıdır:

$$f_X(x) \geq 0$$

$$\sum_x f_X(x) = 1$$



OLASILIK YOĞUNLUK FONKSİYONU (PDF)

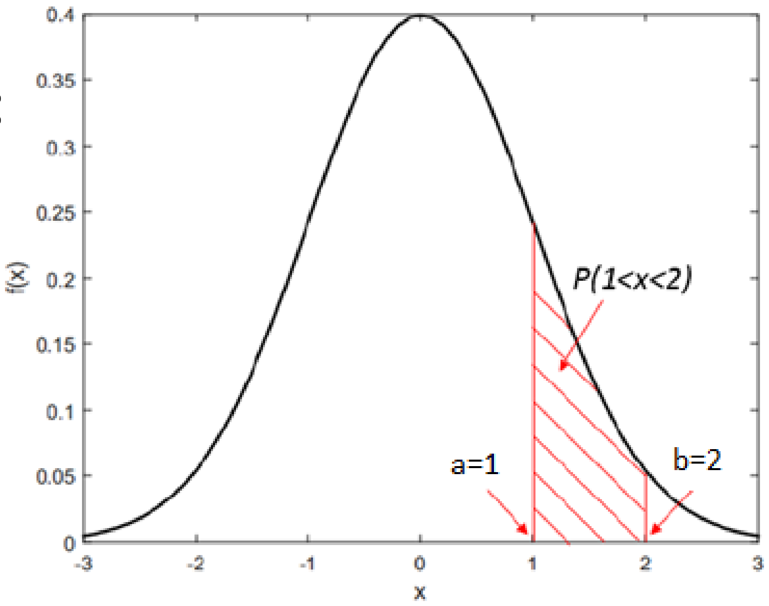
X sürekli bir rasgele değişken olduğunda;

$P(a < X < b) = P(a \leq X \leq b) = \int_a^b f(x)dx$ fonksiyonu rastgele değişkenin olasılık fonksiyonu veya olasılık kütle fonksiyonudur.

Olasılık yoğunluk fonksiyonu aşağıdaki koşulları sağlamalıdır:

$$f_X(x) \geq 0$$

$$\int_{-\infty}^{+\infty} f_X(x)dx = 1$$



BEKLENEN DEĞER (EXPECTED VALUE)

Beklenen değer, ortaya çıkan olasılık dağılımından beklenen değeri sağlar ve her değerin, o değeri elde etme olasılığıyla çarpılması ve olası tüm sonuçların toplanmasıyla hesaplanır:

$$E(X) = \sum_x x f_X(x)$$

$$E(X) = \mu = \sum_i p(x_i) x_i$$

BEKLENEN DEĞER (EXPECTED VALUE)

Sürekli bir değişken için beklenen değeri kesikli bir değişken gibi hesaplayabiliriz, tüm aralık için eğrinin altındaki alanı hesaplamak için aralık boyunca toplamı bir integralle değiştirebiliriz.

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx$$

$$E(X) = \mu = \int_{-\infty}^{+\infty} p(x) x dx$$

BEKLENEN DEĞER (EXPECTED VALUE)

Özetle,

Bir rastgele değişkenin (X) beklenen değeri, E(X), rastgele değişkenin alabileceği değerler kümesinin ortalamasıdır.

Örnek 3: İdeal bir 6 yüzlü bir zarın herhangi bir yüzünün gelme olasılığı 1/6=0,167 olduğuna göre;

Beklenen değeri:

$$E(X) = 1\frac{1}{6} + 2\frac{1}{6} + 3\frac{1}{6} + 4\frac{1}{6} + 5\frac{1}{6} + 6\frac{1}{6} = 3.5$$

BEKLENEN DEĞER (EXPECTED VALUE)

Örnek 4: 1000 kez atılmış 6 yüzlü bir zarın gözlemlerine ait dağılım verildiğine göre;

- Beklenen değeri:

$$E(X) = 1 * 0.195 + 2 * 0.135 + 3 * 0.130 + 4 * 0.192 + 5 * 0.126 + 6 * 0.222 = 3.585$$

VARYANS (VARIANCE)

Belirli bir veri noktası için model tahmininin değişkenliği veya verilerin nasıl yayıldığını bize gösteren değerdir. Analizcinin gözlemlerin ortalama etrafında ne kadar dağınık veya yayılmış olduğunu anlamasını sağlar.

$$E\left((X-\mu)^2\right)=\sigma^2=\sum_i p(x_i)(x_i-\mu)^2$$

Kesikli Rassal Değişken

$$E\left((X-\mu)^2\right)=\sigma^2=\int_{-\infty}^{+\infty} p(x)(x-\mu)^2 dx$$

Sürekli Rassal Değişken

σ: Standart sapma
μ: Ortalama

YANLILIK (BIAS)

Modelleme sonucunda tahmin edilen veriler ile gerçek veriler arasındaki uzaklığı yansıtan değerdir.

EĞİTİM VE TEST VERİ SETİ (TRAIN AND TEST DATA SET)

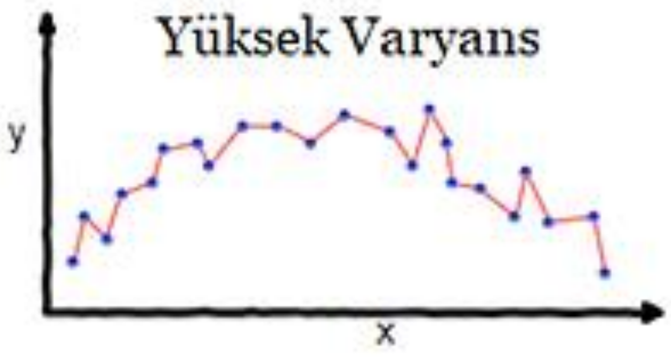
Veri setinde modelin kurulduğu kısmına eğitim veri seti, kurulan modelin doğruluğunun ve nasıl çalıştığını gözlemlenmesi için ayrılan veriye test veri seti denir.

AŞIRI ÖĞRENME (OVERFITTING)

Bir modelin eğitim verisiyle tam uyum göstermeye zorlanması sebebiyle test (veya senaryo) verilerini iyi tahmin edememesi durumudur.

EKSİK ÖĞRENME (UNDERFITTING)

Eğitim veri setinde model eğitilirken sonuçların doğruluğunun düşük çıktığı ve doğal olarak test (veya senaryo) verilerinin doğru tahmin edilemeyeceği durumdur.

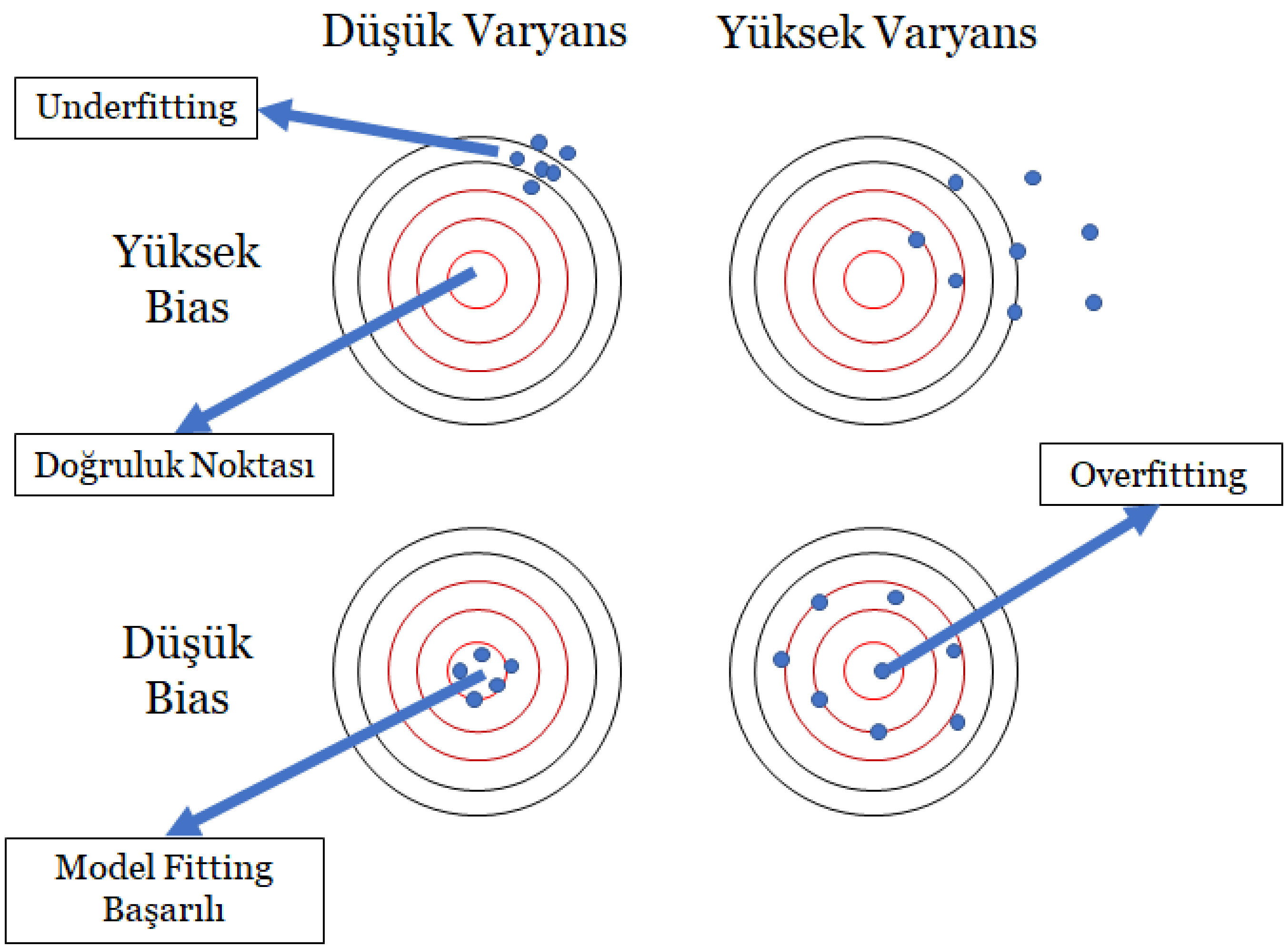


Overfitting

Underfitting

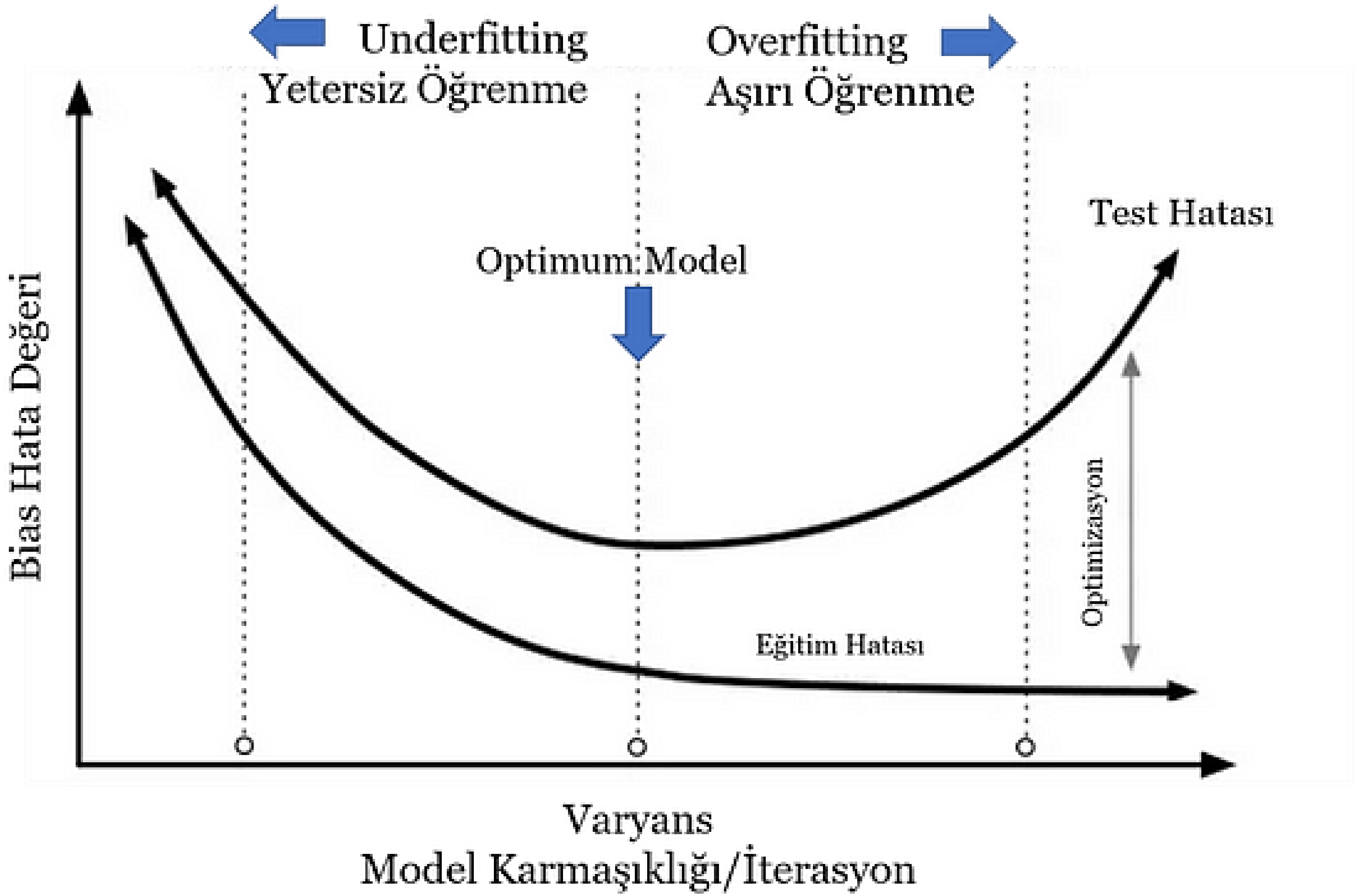
Model Uyumlu
Dengeli Tahmin

VARYANS ve YANLILIK



Bazı Çözüm Yöntemleri:

- Çapraz geçerleme (Cross validation)
- Veri ekleme (Row insertion)
- Değişken seçimi (Feature selection)
- Budama (Trim)
- Düzenlileştirme (Regularization)


Underfitting
Yetersiz Öğrenme
Overfitting
Aşırı Öğrenme
Test Hatası
Optimum Model
Bias Hata Değeri
Optimizasyon
Eğitim Hatası
Varyans
Model Karmaşıklığı/İterasyon

KOVARYANS (COVARIANCE)

İki rastgele değişkenin birlikte değişme derecesini temsil eden istatistiksel bir ölçüdür. Bir kovaryans istatistiğinin işareti, iki rastgele değişken arasındaki ilişkinin yönünü gösterir, ancak büyüklüğü, ilişkinin gücünü temsil etmesi açısından genellikle bilgilendirici değildir.

$$Cov(x_1, x_2) = E((x_1 - \mu_1)\,(x_2 - \mu_2))$$

İki rastgele değişkenin olduğu bir durum için;

$$E(x_1 + x_2) = E(x_1) + E(x_2) = \mu_1 + \mu_2$$

$$Var(x_1 + x_2) = Var(x_1) + Var(x_2) + 2Cov(x_1, x_2)$$

Burada; $Cov(x_1, x_2) = E((x_1 - \mu_1)\,(x_2 - \mu_2))$ olduğuna göre;

$$E(x_1 + x_2) = E(x_1) + E(x_2) = \mu_1 + \mu_2$$

$$Var(x_1 + x_2) = Var(x_1) + Var(x_2) + 2E((x_1 - \mu_1)\,(x_2 - \mu_2))$$

olacaktır.

Örnek 5: Bir zar atılması durumu

X_1	$f(X_1)$	$X_1 f(X_1)$		$(X_1-\mu)^2$	$(X_1-\mu)^2 f(X_1)$
1	0.167	0.167		6.250	1.042
2	0.167	0.333		2.250	0.375
3	0.167	0.500		0.250	0.042
4	0.167	0.667		0.250	0.042
5	0.167	0.833		2.250	0.375
6	0.167	1.000		6.250	1.042
Beklenen Değer, $E(X_1) = (\mu)$		3.500		Varyans, $Var =$	2.917
				Standart Sapma, $\sigma =$	1.708

Örnek 6: İki ayrı zar atılması durumu için kovaryans;

$X_1 X_2 f(X_1, X_2)$

X_1 ve X_2	1	2	3	4	5	6	
1	0.028	0.056	0.083	0.111	0.139	0.167	
2	0.056	0.111	0.167	0.222	0.278	0.333	
3	0.083	0.167	0.250	0.333	0.417	0.500	
4	0.111	0.222	0.333	0.444	0.556	0.667	
5	0.139	0.278	0.417	0.556	0.694	0.833	
6	0.167	0.333	0.500	0.667	0.833	1.000	
							12.250

$$cov(X_1, X_2) = E(X_1 X_2) - E(X_1)E(X_2) = 12.25 - 3.5 * 3.5 = 0$$

VARYANS-KOVARYANS MATRİSİ (VARIANCE-COVARIANCE MATRIX)

$$\sum = \begin{pmatrix} Var(X_1) & cov(X_1, X_2) & cov(X_1, X_3) \\ cov(X_1, X_2) & Var(X_2) & cov(X_2, X_3) \\ cov(X_1, X_3) & cov(X_2, X_3) & Var(X_3) \end{pmatrix}$$

$$cov(X_i, X_i) = Var(X_i)$$

Örnek 6 için; $\sum = \begin{pmatrix} 2.917 & 0 \\ 0 & 2.917 \end{pmatrix}$

KORELASYON (CORRELATION)

Korelasyon katsayısı, sadece bir ilişkinin yönünü belirlemek için değil, aynı zamanda iki rastgele değişken arasında var olabilecek ilişkinin büyüklüğünün gücünü belirlemek için de kullanılır.

$$\rho = \frac{cov(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}}$$

$$-1 \leq \rho \leq +1$$

KORELASYON ve KOVARYANS

Kovaryans, iki veya daha fazla rastgele değişken kümesi arasındaki korelasyonun gücünün bir ölçümüdür. Korelasyon, bir kovaryansın ölçeklendirilmiş hali gibidir.

Kovaryans, -1 ile 1 arasında olmakla sınırlı değildir. Ancak işaret her zaman aynı olacaktır ve sıfır kovaryans ile sıfır korelasyon tam olarak aynı anlama gelir.

Kovaryans değişkenlerin orijinal birimlerinde olduğu için, daha büyük sayılara ve daha geniş dağılımlara sahip ölçeklerdeki değişkenler zorunlu olarak daha büyük kovaryans değerlerine sahip olacaktır.

ORTALAMA (MEAN)

Tüm veri noktalarının toplamının toplam veri noktasına bölümü ile edilen değerdir.

$$\overline{X} = \frac{\sum X}{n}$$

MOD (MODE)

Veri setindeki en sık gözlemlenen değerdir.

MEDYAN (MEDIAN)

Sıralı bir veri setindeki orta değeri göstermektedir.

ÖRNEKLEM İSTATİSTİKLERİ

Örneklem varyansı: $S_x{}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$

Örneklem kovaryansı: $cov(X_1, X_2) = \frac{\sum(X_1 - \bar{X}_1)(X_2 - \bar{X}_2)}{n-1}$

Örneklem korelasyon katsayısı: $r = \frac{cov(X_1, X_2)}{S_{x_1} S_{x_2}}$

KESİKLİ TERCİH MODELLERİ
Discrete Choice Models

TEŞEKKÜRLER
HAFTAYA GÖRÜŞMEK ÜZERE ☺

Dr. Kadir Berkhan AKALIN

KAYNAK GÖSTERME

Bu sunuma aşağıdaki gibi atıf yapabilirsiniz:
Akalın, K.B. (2023). Kesikli Tercih Modelleri Ders Notu. Eskişehir Osmangazi Üniversitesi Fen Bilimleri Enstitüsü

You can cite this presentation as follows:
Akalın, K.B. (2023). Discrete Choice Models Lecture Notes. Eskisehir Osmangazi University Graduate School of Natural and Applied Sciences.

KAYNAKLAR

- *Akalın, K.B. (2021). Yolculuk Üretim ve Çekim Modellerinin Rastgele Pişmanlık Minimizasyonu ve Rastgele Fayda Maksimizasyonu Yöntemleri ile Geliştirilmesi. Doktora Tezi. Eskişehir Osmangazi Üniversitesi.*
- *Ben-Akiva, M., Bierlaire, M. (1999). Discrete choice methods and their applications to short term travel decisions. Handbook of transportation science.*
- *Ben-Akiva, M., Lerman, S. (1985). Discrete Choice Analysis, The MIT Press.*
- *De Dios Ortúzar, J., Willumsen, L.G. (2011). Modelling Transport. John Wiley & Sons.*
- *Frumin, M., Ben-Akiva M. (2008). Transportation Systems Analysis: Demand And Economics. MIT Open Courseware.*
- *Tezcan, H.O. (2021). Discrete Choice Modelling in Transportation Lecture Notes. İstanbul Technical University.*
- *Tezcan, H.O. (2021). Transportation Models Lecture Notes. İstanbul Technical University.*
- *Train, K. (2002). Discrete Choice Methods with Simulation, Cambridge University Press.*
- https://tr.wikipedia.org/wiki/Olas%C4%B1l%C4%B1k_da%C4%9F%C4%B1l%C4%B1m%C4%B1
- https://www.wasyresearch.com/continuous-and-discrete-statistical-distributions-probability-density-mass-function-cumulative-distribution-function-and-the-central-limit-theorem/
- https://acikders.ankara.edu.tr/pluginfile.php/793/mod_resource/content/2/Olas%C4%B1l%C4%B1k%20ve%20Olas%C4%B1l%C4%B1k%20Da%C4%9F%C4%B1l%C4%B1mlar%C4%B1.pdf
- https://towardsdatascience.com/probability-mass-and-density-functions-eab86a81d021
- https://towardsdatascience.com/bias-variance-and-how-they-are-related-to-underfitting-overfitting-4809aed98b79
- https://medium.com/datarunner/bias-ve-varyans-dengesi-%C3%BCzerinden-overfitting-ve-underfitting-durumlar%C4%B1n%C4%B1n-anla%C5%9F%C4%B1lmas%C4%B1-ca26cea65bb4