Model-veri ilişkisi

Sinan Yıldırım

MDBF, Sabancı Üniversitesi

7 Şubat 2023

Bu sunumdaki şekiller, şu kitaplardan alınmıştır.

Pattern Recognition and Machine Learning, Christopher M. Bishop;

Mathematics for Machine Learning, Deisenroth v.d.

Hazırlık için: Olasılık ve dağılımlar

Olasılık uzayı $(\Omega, \mathcal{A}, P)$

- **Örneklem uzayı** Ω

- **Olay uzayı** $\mathcal{A}$

  Olay $A \in \mathcal{A}$

- **Olasılık** $P : \mathcal{A} \mapsto [0, 1]$.

Olasılık uzayı $(\Omega, \mathcal{A}, P)$

Rassal değişken

- **Hedef uzay** $\mathcal{T}$ (çoğu zaman gerçel sayılardan oluşur.)
- **Rassal değişken**

$$X : \Omega \mapsto \mathcal{T}$$

- Bir $S \subseteq \mathcal{T}$ için

$$P_X(X \in S) = P(X^{-1}(S)) = P \circ X^{-1}(S)$$

X'in **olasılık dağılımı (kanunu)**: $P \circ X^{-1}$.

  - Ayrık rassal değişken: $\mathcal{T}$ sayılabilir.
  - Sürekli rassal değişken: Örnek $\mathcal{T} = \mathbb{R}$.

Kümülatif dağılım fonsksiyonu

- Tek boyutlu değişken:

$$F(x) := P(X \leq x), \quad x \in \mathbb{R}.$$

- Çok boyutlu değişken: $X = (X_1, \ldots, X_D)$.
  $\boldsymbol{x} = (x_1, \ldots, x_D)$ için,

$$F(\boldsymbol{x}) := P(X_1 \leq x_1, \ldots, X_D \leq x_D), \quad \boldsymbol{x} \in \mathbb{R}^D.$$

Ayrık dağılımlar

$\mathcal{T}$ sayılabilir.

- Tek boyutlu değişkenler:
  Olasılık kütle fonksiyonu:

$$p(x) = P(X = x), \quad x \in \mathcal{T}$$

- Çok boyutlu değişkenler:
  $\mathcal{T}$ kartezyen çarpımı
  Bileşik olasılık kütle fonksiyonu (örn. iki değişken için)

$$p(x, y) = P(X = x, Y = y), \quad (x, y) \in \mathcal{T}$$

Sürekli dağılımlar

Olasılık yoğunluk fonksiyonu: $f : \mathbb{R}^D \mapsto [0, \infty)$

X'in olasılık yoğunluk fonksiyonu f ise,

Bir aralığın olasılığı:

$P(a \leq X \leq b)$

Bir değerin olasılığı:

$P(X = a)$

Toplam kuralı

Bileşik dağılım: $p(\boldsymbol{x}, \boldsymbol{y})$

Marjinal dağılım:

- $p(\boldsymbol{x})$
- $p(\boldsymbol{y})$

Koşullu olasılık ve çarpım kuralı

Olasılık uzayı: $(\Omega, \mathcal{A}, P)$

$A, B \in \mathcal{A}$ kümeleri için,

$P(B|A)$: B olayının A olayına **koşullu olasılığı**:

$$P(B|A) := \frac{P(A \cap B)}{P(A)}$$

Çarpım kuralı:

$$P(A \cap B) = P(B|A)P(A)$$

Koşullu olasılık ve çarpım kuralı

Dağılımlar için:

$p(\boldsymbol{y}|\boldsymbol{x})$: Y'nin $X = \boldsymbol{x}$'e koşullu dağılımı

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{y})}$$

Çarpım kuralı:

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x})$$

Örnek

$p(x,y)$	$y=1$	$y=2$	$y=3$	$p(x)$
$x=1$	0.0	0.1	0.2	
$x=2$	0.4	0.2	0.1	
$p(y)$				

$p(x\|y)$	$y=1$	$y=2$	$y=3$
$x=1$			
$x=2$			

Bayes teoremi

Çarpım kuralının basit bir uygulaması:

Bayesci istatistik

Alıştırma

Bir suç mahalinde DNA izine rastlanıyor. Şüphelilerden DNA örneği alınıp bulunan DNA ile karşılaştırılıyor.

Testin performansı:

	Pozitif	Negatif
Suçlu	1	0
Suçsuz	0.001	0.999

Suçlu olma öncül olasılığı 10^{-5} olan bir A kişisi teste tabi tutuluyor ve testin sonucu pozitif çıkıyor.
A kişisi suçlu mudur? A kişisinin testin sonucuna koşullu suçlu olma olasılığını bulun.

Beklenti (Ortalama)

Beklenti doğrusal bir operatördür:

Ortalama, Medyan, Doruk

Varyans (tek boyutta)

Kovaryans

- Tek boyutta
- Çok boyutta

Kovaryans matrisi

Korelasyon

Ampirik ortalama

- Tek boyutta

$$\bar{x} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

- Çok boyutta

Ampirik varyans/kovaryans matrisi

- Tek boyutta

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_i - \bar{x})^2$$

- Çok boyutta

Toplamlar ve dönüşümler

- $\mathbb{E}(\boldsymbol{x} + \boldsymbol{y})$
- $\mathbb{E}(\boldsymbol{x} - \boldsymbol{y})$
- $\mathbb{V}(\boldsymbol{x} + \boldsymbol{y})$
- $\mathbb{V}(\boldsymbol{x} - \boldsymbol{y})$

Afin dönüşümler

$\boldsymbol{x}$: ortalaması $\boldsymbol{\mu}$ ve kovaryans matrisi $\boldsymbol{\Sigma}$

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$$

- $\mathbb{E}(\boldsymbol{y})$
- $\mathbb{V}(\boldsymbol{y})$
- $\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y})$

Bağımsızlık

X ve Y'nin bağımsızlığı için gerekli ve yeterli koşul

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x})p(\boldsymbol{y}), \quad \forall \boldsymbol{x}, \boldsymbol{y}$$

Gösterim: $X \perp\!\!\!\perp Y$

X ve Y bağımsız ise,

- $p(\boldsymbol{y}|\boldsymbol{x})$
- $p(\boldsymbol{x}|\boldsymbol{y})$
- $\mathbb{V}_{X,Y}[\boldsymbol{x} + \boldsymbol{y}]$
- $\mathrm{Cov}_{X,Y}[\boldsymbol{x}, \boldsymbol{y}]$

Koşullu bağımsızlık

X ve Y'nin Z'ye **koşullu bağımsızlığı** için gerekli ve yeterli koşul

$$p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{z}), \quad \forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$$

Gösterim: $X \perp\!\!\!\perp Y \mid Z$

Gauss dağılımı (Normal dağılım)

Tek boyutlu değişken: $X \sim \mathcal{N}(\mu, \sigma^2)$.

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad x \in \mathbb{R}$$

Yoğunluk fonksiyonu $\mathcal{N}(x|\mu, \sigma^2)$ olarak da gösterilir.

Gauss dağılımı (Normal dağılım)

Çok boyutlu değişken: $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{|2\pi\boldsymbol{\Sigma}|^{-1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right\}, \quad \boldsymbol{x} \in \mathbb{R}^D$$

Yoğunluk fonksiyonu $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ olarak da gösterilir.

Gauss dağılımı: marjinal ve koşullu dağılımlar

$$p(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} \middle| \begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{bmatrix} \right)$$

- $p(\boldsymbol{x})$
- $p(\boldsymbol{y})$
- $p(\boldsymbol{x}|\boldsymbol{y})$

Gauss dağılımı: toplam

$X \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$, $Y \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$; X, Y bağımsız ve aynı boyda.

$Z = aX + bY$

- $\mathbb{E}[\boldsymbol{z}]$
- $\mathbb{V}[\boldsymbol{z}]$
- $p(\boldsymbol{z})$

Gauss dağılımı: doğrusal dönüşüm

$X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

$\boldsymbol{A} \in \mathbb{R}^{M \times N}$, $Y = \boldsymbol{A}X$.

Y'nin dağılımı?

Gauss dağılımı: doğrusal dönüşüm

$Y \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{A} \in \mathbb{R}^{M \times N}$

$Y = \boldsymbol{A}X.$

X'in dağılımı?

- $M = N$, $\boldsymbol{A}$ terslenebilir:
- $M > N$, $\boldsymbol{A}$'nın kertesi N:

Alıştırma

$X = (X_1, \ldots, X_D)^T \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ise, $Y = X_1 + \ldots + X_D$'nin dağılımını bulun.

Gauss dağılımı: karışım

$p_1(x) = \mathcal{N}(x, \mu_1, \sigma_1^2)$, $p_2(x) = \mathcal{N}(x, \mu_2, \sigma_2^2)$ ve $0 < \alpha < 1$ olsun.

Bir karışım dağılımı:

$$p(x) = \alpha p_1(x) + (1 - \alpha) p_2(x)$$

- $\mathbb{E}(x)$
- $\mathbb{V}(x)$

Alıştırma: Bu karışım dağılımı bir Gauss dağılımı mıdır?

Gauss dağılımı: örnekleme

Bernoulli dağılımı

Başarı olasılığı $\mu \in [0, 1]$

$$p(x|\mu) = \mu^x (1-\mu)^{1-x}, \quad x = 0, 1.$$

- $\mathbb{E}[x]$
- $\mathbb{V}[x]$

Binom dağılımı

N bağımsız deneme, her bir denemede başarı olasılığı $\mu \in [0, 1]$

$$p(x|N, \mu) = \binom{N}{x} \mu^x (1 - \mu)^{N-x}, \quad x = 0, \ldots, N.$$

- $\mathbb{E}[x]$
- $\mathbb{V}[x]$

Beta dağılımı

Rassal değişken: $\mu \in [0, 1]$,
Parametreler: $\alpha > 0$, $\beta > 0$

$$p(\mu|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1}(1-\mu)^{\beta-1}, \quad \mu \in [0, 1].$$

- $\mathbb{E}[\mu]$
- $\mathbb{V}[\mu]$

Eşleniklik

Bayes teoremi:

$$\text{sonsal dağılım} = \frac{\text{öncül dağılım} \times \text{olabilirlik}}{\text{kanıt}}$$
$$\propto \text{öncül dağılım} \times \text{olabilirlik}$$

Eşlenik dağılım: Öncül dağılım ile sonsal dağılım aynı forma sahipse, o öncül dağılım olabilirlik fonksiyonu için eşlenik bir dağılımdır.

Beta-Binom eşlenikliği

$$p(x|N,\mu) = \binom{N}{x} \mu^x (1-\mu)^{N-x}, \quad x = 0, \ldots, N.$$

$$p(\mu|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1}(1-\mu)^{\beta-1}, \quad \mu \in [0,1].$$

Yeterli istatistik

Bir veya birden fazla rassal değişkenin herhangi bir belirlenimci fonksiyonuna istatistik denir.

Yeterli istatistik: X'in dağılımı $p(x|\theta)$ olsun. Bir $\phi(x)$ istatistiğinin θ için yeterli istatistik olması için yeterli ve gerekli koşul:

$$p(x|\theta) = h(x)g_\theta(\phi(x))$$

$h(x) \geq 0$: θ'dan bağımsız bir fonksiyon.

Hangi dağılımlarda (x'in kendisinden başka) yeterli istatistiğe rastlanabilir?

Üstel dağılım aileleri

Bir üstel dağılım ailesinin üyeleri $\boldsymbol{\theta} \in \mathbb{R}^D$ ile şu şekilde belirlenir

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = h(\boldsymbol{x}) \exp(\boldsymbol{\theta}^T \phi(\boldsymbol{x}) - A(\boldsymbol{\theta}))$$

- $\phi(\boldsymbol{x}) \in \mathbb{R}^D$: yeterli istatistikler
- $\boldsymbol{\theta}$: doğal parametre
- $A(\boldsymbol{\theta})$: düzgeleştirici (normalize edici) katsayı (log-bölüntüleme fonksiyonu)
- $h(\boldsymbol{x}) \geq 0$: $\boldsymbol{\theta}$'dan bağımsız bir fonksiyon.

Gauss dağılımı

Bernoulli dağılımı

Alıştırma

Parametresi $\lambda \in [0, \infty)$ olan Poisson dağılım ailesini ele alalım.

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, \ldots$$

Bu aile bir üstel dağılım ailesi midir?

Alıştırma

Parametreleri $a, b \in \mathbb{R}$, $a < b$ olan homojen dağılım ailesini ele alalım:

$$p(x|a,b) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{değilse} \end{cases}$$

Bu aile bir üstel dağılım ailesi midir?

Üstel dağılımlar ve eşleniklik

Olabilirlik fonksiyonu üstel dağılım ise, mutlaka eşlenik dağılımı vardır.

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = h(\boldsymbol{x}) \exp\left(\boldsymbol{\theta}^T \phi(\boldsymbol{x}) - A(\boldsymbol{\theta})\right)$$

olabilirlik fonksiyonu için,

$$p(\boldsymbol{\theta}|\gamma) = h_c(\boldsymbol{\theta}) \exp\left(\left[\boldsymbol{\gamma}_1^T \boldsymbol{\theta} - \gamma_2 A(\boldsymbol{\theta})\right] - A_c(\gamma)\boldsymbol{\theta}\right)$$

öncül dağılımı eşleniktir.

Öncül dağılım için:

- Doğal parametre $\gamma = \begin{bmatrix} \boldsymbol{\gamma}_1 \\ \gamma_2 \end{bmatrix}$
- Yeterli istatistik: $\begin{bmatrix} \boldsymbol{\theta} \\ -A(\boldsymbol{\theta}) \end{bmatrix}$.

Bernoulli dağılımı için eşlenik öncül

Alıştırma

Üstel dağılım ailesinden bir

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = h(\boldsymbol{x}) \exp\left(\boldsymbol{\theta}^T \phi(\boldsymbol{x}) - A(\boldsymbol{\theta})\right)$$

olabilirlik fonksiyonu için, her zaman eşlenik bir öncül dağılımın

$$p(\boldsymbol{\theta}|\gamma) = h_c(\boldsymbol{\theta}) \exp\left(\left[\gamma_1^T \boldsymbol{\theta} - \gamma_2 A(\boldsymbol{\theta})\right] - A_c(\gamma)\right)$$

şeklinde önerilebilir. Bu öncül dağılım ve olabilirlikten yola çıkarak sonsal dağılımı türetip sonsal dağılımın öncül dağılımla aynı yapıda olduğunu gösterin. (Sonsal dağılımın doğal parametrelerini (γ'larını) belirleyin.)

Değişken dönüşümü

Diyelim ki X'in dağılımını biliyoruz.
$Y = U(X)$.
Y'nin dağılımı nedir?

Değişken dönüşümü - ayrık dağılımlar

$$Y = U(X)$$

U tersi alınabilir bir fonksiyon ise

$$\begin{aligned} P(Y = y) &= P(U(X) = y) \\ &= P(X = U^{-1}(y)). \end{aligned}$$

Değişken dönüşümü - sürekli dağılımlar

$$Y = U(X)$$

Kümülatif dağılım fonksiyonu tekniği:

$$F_Y(y) = P(Y \leq y)$$

Örnek

X'in dağılımı

$$f_X(x) = 3x^2, \quad x \in [0, 1].$$

$Y = X^2$'nin dağılımı nedir?

Değişken dönüşümü - genel

U tersi alınabilir ve artan ise

$$
\begin{aligned}
F_Y(y) &= P(U(X) \leq y) \\
&\leq P(X \leq U^{-1}(y)) \\
&= \int_{-\infty}^{U^{-1}(y)} f(x)dx
\end{aligned}
$$

İfadenin y'ye göre türevi Y'nin olasılık dağılım fonksiyonunu verir.

U tersi alınabilir ve azalan olsaydı da aynı ifadeyi elde edecektik.

Değişken dönüşümü - çok boyutlu

$X \in \mathbb{R}^D$

$Y = U(X).$

$\boldsymbol{y} = U(\boldsymbol{x})$ tersi alınabilir ve türevlenebilir ise

$$f_Y(\boldsymbol{y}) = f_X(U^{-1}(\boldsymbol{y})) \left| \det \left(\frac{\partial}{\partial \boldsymbol{y}} U^{-1}(\boldsymbol{y}) \right) \right| .$$

Alıştırma

$X \sim \mathcal{N}(0,1)$ ise değişken dönüşümü tekniklerinden uygun birini kullanarak $Y = \exp(X)$'in dağılımını bulun.

Model-veri ilişkisi

Örnek-etiket tipi veriler

Veri:

$$(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$$

- $\boldsymbol{x}_n$: örnek
- y_n: etiket

Amaçlar

- Veriyi açıklama
- Yeni bir $\boldsymbol{x}$ verildiğinde y'yi tahminleme.
- vb

Örnek veri

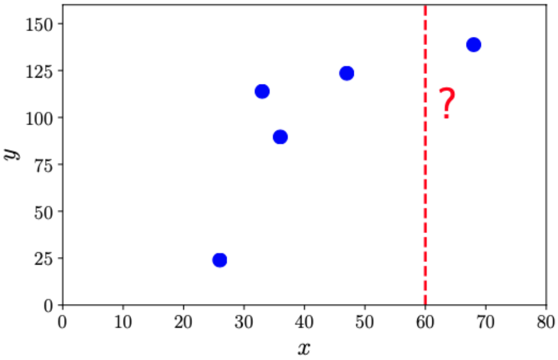
Name	Gender	Degree	Postcode	Age	Annual salary
Aditya	M	MSc	W21BG	36	89563
Bob	M	PhD	EC1A1BA	47	123543
Chloé	F	BEcon	SW1A1BH	26	23989
Daisuke	M	BSc	SE207AT	68	138769
Elisabeth	F	MBA	SE10AA	33	113888

Örnek veri - sayısallaştırılmış

Gender ID	Degree	Latitude (in degrees)	Longitude (in degrees)	Age	Annual Salary (in thousands)
-1	2	51.5073	0.1290	36	89.563
-1	3	51.5074	0.1275	47	123.543
+1	1	51.5071	0.1278	26	23.989
-1	1	51.5075	0.1281	68	138.769
+1	2	51.5074	0.1278	33	113.888

Tahminleme

x: yaş,
y: maaş



Tahminleyici

Aday fonksiyonlar:

$$f(\cdot, \boldsymbol{\theta}) : \mathbb{R}^d \mapsto \mathbb{R}.$$

$\boldsymbol{\theta}$: Tahminleyicinin parametresi

Amaç:

$$f(\boldsymbol{x}_n, \boldsymbol{\theta}^*) \approx y_n, \quad \forall n = 1, \ldots, N.$$

olacak şekilde bir $\boldsymbol{\theta}^*$ belirlemek.

Tahmin: $\hat{y}_n = f(\boldsymbol{x}_n, \boldsymbol{\theta}^*)$.

Örnek: Doğrusal regresyon

$$\boldsymbol{x}_n = \begin{bmatrix} 1 & x_n^{(1)} & x_n^{(2)} & \dots & x_n^{(D)} \end{bmatrix}^T, \quad y \in \mathbb{R}$$

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_0 & \theta_1 & \dots & \theta_D \end{bmatrix}^T$$

$f(\cdot, \boldsymbol{\theta}) : \mathbb{R}^{D+1} \mapsto \mathbb{R}$,

$$f(\boldsymbol{x}_n, \boldsymbol{\theta}) = \boldsymbol{\theta}^T \boldsymbol{x}_n$$

Ampirik risk enküçültme

Ampirik risk

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_N \end{bmatrix}^T \in \mathbb{R}^{N \times D}, \quad \boldsymbol{y} = \begin{bmatrix} y_1 & \dots & y_N \end{bmatrix}^T$$

Tahmin

$$\hat{y}_n = f(\boldsymbol{x}_n, \boldsymbol{\theta}).$$

Kayıp

$$\ell(y_n, \hat{y}_n).$$

Ampirik risk:

$$\boldsymbol{R}_{\text{amp}}(f, \boldsymbol{X}, \boldsymbol{y}) = \frac{1}{N} \sum_{n=1}^{N} \ell(y_n, \hat{y}_n).$$

Ampirik risk enküçültmesi:

$$f^* = \arg\min_f \boldsymbol{R}_{\text{emp}}(f, \boldsymbol{X}, \boldsymbol{y})$$

Örnek: En küçük kareler

Karesel kayıp:

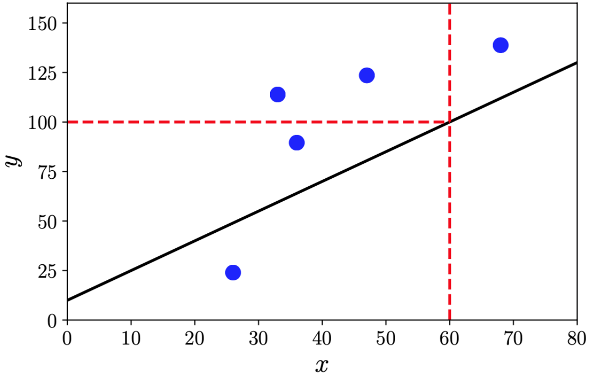
$$\ell(y_n, \hat{y}_n) = (y_n - \hat{y}_n)^2$$

Doğrusal tahminleyici:

$$f(\boldsymbol{x}_n, \boldsymbol{\theta}) = \boldsymbol{\theta}^T \boldsymbol{x}_n.$$

Ampirik risk enküçültmesi:

$$\begin{aligned}
\boldsymbol{\theta}^* &= \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^D} \frac{1}{N} \sum_{n=1}^{N} (y_n - \boldsymbol{\theta}^T \boldsymbol{x}_n)^2 \\
&= \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^D} \frac{1}{N} \left\| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta} \right\|^2.
\end{aligned}$$


150
125
100
75
50
25
0
y
0
10
20
30
40
50
60
70
80
x

Gerçek ve ampirik risk

Gerçek risk

$$\boldsymbol{R}_{\text{ideal}}(f) = \mathbb{E}_{\boldsymbol{x},y}[\ell(y, f(\boldsymbol{x}))]$$

Ampirik risk, gerçek riski kestirir.

$$\boldsymbol{R}_{\text{amp}}(f, \boldsymbol{X}, \boldsymbol{y}) \approx \boldsymbol{R}_{\text{ideal}}(f).$$

Aşırı uyma problemi

Eğitim sonucu, *görünmeyen veriye* ne kadar uyumlu?

Bunu kestirmek için genelde veri ikiye bölünür:

$$(\boldsymbol{X}_{\text{train}}, y_{\text{train}}), \quad (\boldsymbol{X}_{\text{test}}, y_{\text{test}})$$

İstenen:

$$R_{\text{amp}}(f, \boldsymbol{X}_{\text{train}}, y_{\text{train}}) \approx R_{\text{amp}}(f, \boldsymbol{X}_{\text{test}}, y_{\text{test}})$$

Aşırı uyma:

$$R_{\text{amp}}(f, \boldsymbol{X}_{\text{train}}, y_{\text{train}}) < R_{\text{amp}}(f, \boldsymbol{X}_{\text{test}}, y_{\text{test}})$$

Düzenleme

Aşırı uyumun belirtilerinden biri $\boldsymbol{\theta}$ bileşenlerinin yüksek değerler almasıdır.

Düzenlenmiş problem

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^D} \frac{1}{N} \|y - \boldsymbol{X}\boldsymbol{\theta}\|^2 + \lambda \|\boldsymbol{\theta}\|^2 .$$

- $\|\boldsymbol{\theta}\|^2$: düzenleyici
- λ: düzenleme katsayısı
- $\lambda \|\boldsymbol{\theta}\|^2$: ceza

Genelleme başarımı

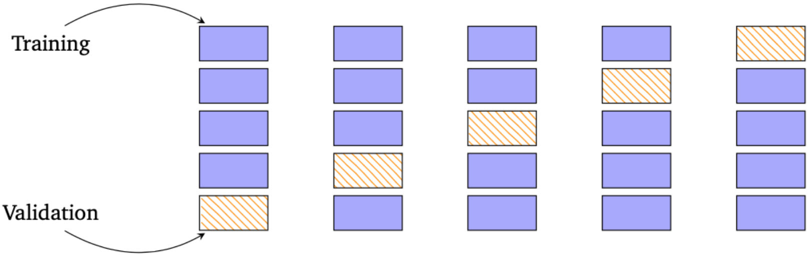
Belli bir veri üzerinde eğitilen bir model aynı kaynaktan başka verileri iyi tahminler mi?

Seçilmiş olan aday fonksiyon ailesine, yani modele dair bir ölçüt.

Belli bir $\boldsymbol{\theta}$'yı değil, $\{f(\cdot, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ ailesini ilgilendiren bir ölçüt.

Genelleme başarımı nasıl ölçülür?

K katlı çapraz doğrulama



$K-1$ parçada eğit, kalan parçada sına. K kez tekrarla.

$$\mathbb{E}_{\mathcal{V}}[R(f,\mathcal{V})] \approx \frac{1}{K}\sum_{k=1}^{K} R(f^{(k)},\mathcal{V}^{(k)})$$

Parametre kestirimi

Parametre kestirimi

Kayıp fonksiyonları yerine olasılık dağılımları kullanılır.

$\boldsymbol{\theta}$: veriyi açıklamak için kullanılan olasılık dağılımının parametresi.

Parametre kestirimi: En iyi $\boldsymbol{\theta}$'yı bulmak.

Örnek - etiket veri modeli

Bağımsız özdeş ikililer:

$$(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$$

$$\mathcal{Y} = \{y_1, \ldots, y_N\}, \quad \mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}.$$

Genelde $\boldsymbol{x}_n$ rassal kabul edilmez:

$$p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta}) = \prod_{n=1}^{N} p(y_n|\boldsymbol{x}_n, \boldsymbol{\theta}).$$

Olabilirlik fonksiyonu

Negatif log-olabilirlik

$$\mathcal{L}(\boldsymbol{\theta}) = -\log p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta}) = -\sum_{n=1}^{N} \log p(y_n|\boldsymbol{x}_n, \boldsymbol{\theta}).$$

Enbüyük olabilirlik kestirimi:

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$$

Gauss dağılımı ve en küçük kareler

$$p(y_n|\boldsymbol{x}_n, \boldsymbol{\theta}) = \mathcal{N}(y_n|\boldsymbol{x}_n^T\boldsymbol{\theta}, \sigma^2).$$

Negatif log-olabilirlik

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \boldsymbol{x}_n^T\boldsymbol{\theta})^2 - \sum_{n=1}^{N}\log\frac{1}{\sqrt{2\pi\sigma^2}}.$$

Enbüyük sonsal dağılım kestirimi

Veri $\mathcal{D}$, parametre: $\boldsymbol{\theta}$

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\boldsymbol{\theta})p(\mathcal{D}|\boldsymbol{\theta})}{p(\mathcal{D})} \propto p(\boldsymbol{\theta})p(\mathcal{D}|\boldsymbol{\theta}).$$

$$\theta^* = \arg\max_{\theta} p(\boldsymbol{\theta}|\mathcal{D})$$

Önceki örneklerde, $\mathcal{D} = (\mathcal{X}, \mathcal{Y})$.

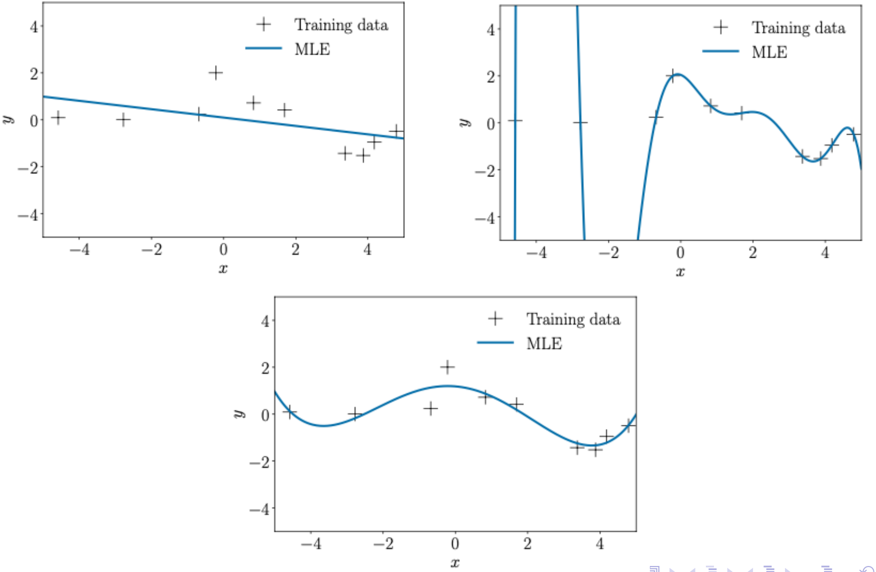
Hedef fonksiyon:

$$p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta})p(\boldsymbol{\theta})$$

($\mathcal{X}$'in olasılık dağılımı olsa da aynı hedef kullanılabilir.)

Aşırı uyum, yetersiz uyum, kararında uyum

Olasılık modeli (görünmeyen) veri için ne kadar uygun?



Olasılıksal modelleme ve çıkarım

Olasılıksal modelleme

Hem parametre hem de veri rassal değişken:

- $\boldsymbol{x}$: gözlemlenen değişken (veri)
- $\boldsymbol{\theta}$ bilinmeyen değişken, parametre

Çıkış noktası: Ortak dağılım

$$p(\boldsymbol{x}, \boldsymbol{\theta})$$

Ortak dağılımın içerdiği bilgiler:

- Öncül dağılım $p(\boldsymbol{\theta})$ ve olabilirlik $p(\boldsymbol{x}|\boldsymbol{\theta})$.
- Marjinal olabilirlik $p(\boldsymbol{x})$ (model seçimi için gerekli)
- Sonsal dağılım $p(\boldsymbol{\theta}|\boldsymbol{x})$

Bayesci çıkarım

Hedef: Sonsal dağılımın belirlenmesi.

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})}{p(\boldsymbol{x})}$$

Tahminleme:

$$p(\boldsymbol{x}_{yeni}|\boldsymbol{x}) = \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{x})p(\boldsymbol{x}_{yeni}|\boldsymbol{\theta}, \boldsymbol{x})d\boldsymbol{\theta}$$

Noktasal kestirimler ile tahminleme:

$$p(\boldsymbol{x}_{yeni}|\boldsymbol{x}) \approx p(\boldsymbol{x}_{yeni}|\boldsymbol{\theta}^*, \boldsymbol{x})$$

Saklı değişkenli modeller

z: saklı değişkeni

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \int p(\boldsymbol{x}|z, \boldsymbol{\theta}) p(z) dz$$

Ortak dağılım:

$$p(\boldsymbol{\theta}, \boldsymbol{x}, z) = p(\boldsymbol{\theta}) p(z) p(\boldsymbol{x}|\boldsymbol{\theta}, z)$$

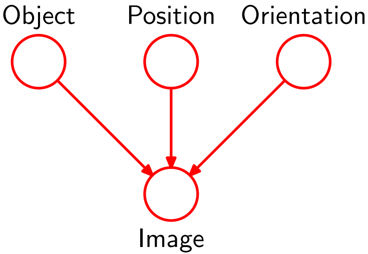
Kullanımı?

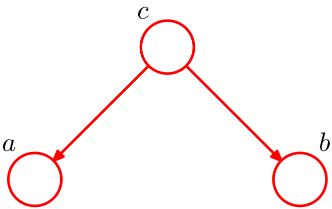
$$p(\boldsymbol{\theta}|z, \boldsymbol{x}), \quad p(z|\boldsymbol{\theta}, \boldsymbol{x})$$

Çizge modelleri

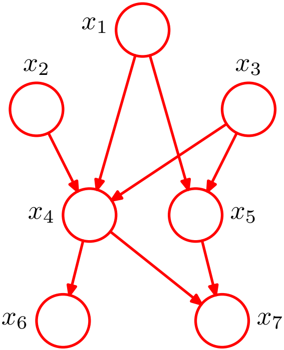
Yönlü düz çizgeler (Bayes ağları)

Rassal değişkenlerin arasındaki koşullu bağımsızlık ilişkilerini sebep sonuç ilişkisini temel alarak verir.





Yönlü düz çizgeler - Olasılık dağılımı



$$p(x_{1:7}) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$$

Genel:

$$p(\boldsymbol{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{Pa}_k)$$

Örnek

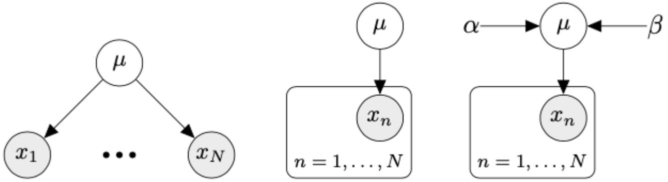
Bir madeni para N kere atılıyor. Yazı: $x = 1$, Tura: $x = 0$

$$p(x|\mu) = \text{Ber}(x|\mu)$$

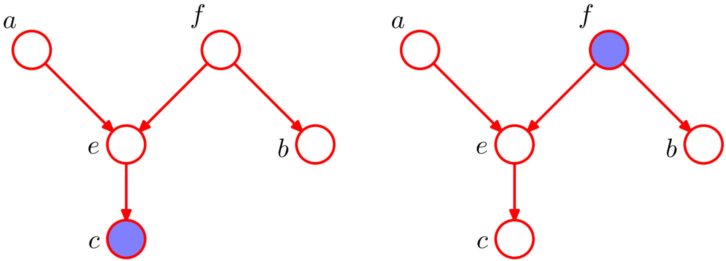
$$p(x_1, \ldots, x_N|\mu) = \prod_{n=1}^{N} p(x_n|\mu)$$

Öncül dağılım $p(\mu) = \text{Beta}(\mu|\alpha, \beta)$.

Üç farklı gösterim



Koşullu bağımsızlık



- $a \perp\!\!\!\perp b | c$ mıdır?
- $a \perp\!\!\!\perp b | f$ mıdır?

Koşullu bağımsızlık - *d*-ayrılık

A, B ve C düğüm kümeleri olsun.
A ve B C'ye koşullu bağımsız mıdır?

$$A \perp\!\!\!\perp B | C$$

d-ayrılık

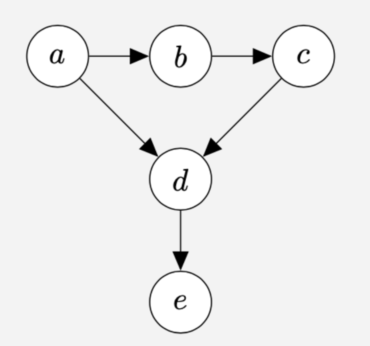
A'daki bir düğümden dan B'deki bir düğüme giden bir yol ele alalım.
Bu yolun üstündeki herhangi bir düğümü ele alalım:

- Bu düğüme ulaşan oklar **kuyruk-kuyruğa** ve **kafa-kuyruğa** ise ve bu düğüm C'nin içindeyse; veya
- Bu düğüme gelen oklar kafa-kafaya ise ve ne bu düğüm ne de onun alt-düğümleri C'nin içinde ise (neither-nor);

bu yol C tarafından engellenmiş sayılır.
Eğer A'daki her bir düğümden dan B'deki her bir düğüme giden bütün yollar C tarafından engellenmişse, A ve B, C tarafından d-ayrılmıştır, ve

$$A \perp\!\!\!\perp B | C$$

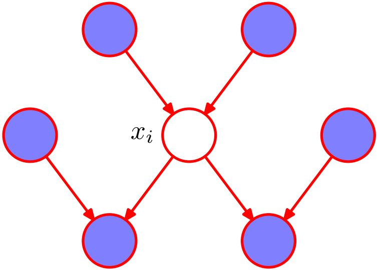
sağlanır.



$$b \perp\!\!\!\perp d \mid a, c \quad ?$$
$$a \perp\!\!\!\perp c \mid e \quad ?$$
$$b \perp\!\!\!\perp d \mid c \quad ?$$
$$a \perp\!\!\!\perp c \mid b, e \quad ?$$

Markov battaniyesi

- Bir üst-düğümler (ebeveynler)
- bir alt-düğümler (çocuklar),
- çocukarın beraber yapıldığı partnerler



Bir düğüm, Markov battaniyesi verildiğinde diğer düğümlerden bağımsızdır.

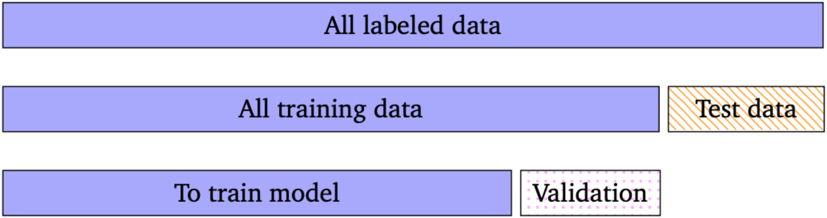
Model seçimi

Model seçimi problemleri

Örnekler:

- Bir örneklem için popülasyon dağılımı seçimi
- Regresyon için kullanılan polinomun derecesi seçimi
- Regresyonda açıklayıcı değişkenlerin seçimi
- Karışım dağılımlarında bileşen sayısı seçimi
- Temel bileşen analizinde boyut seçimi
- Destek vektör makinelereinde çekirdek seçimi
- Derin öğrenmede ağ yapısının seçimi

İç içe geçmiş çapraz doğrulama



$$\mathbb{E}_{\mathcal{V}}[R(\mathcal{V}|M)] \approx \frac{1}{K}\sum_{k=1}^{K} R(\mathcal{V}^{(k)}|M)$$

Tüm M modelleri için hesapla ve en iyisini seç.

Bayesci model seçimi

Modelin kendisi de bir değişken:

$$M \sim p(M)$$
$$\boldsymbol{\theta}|M \sim p(\boldsymbol{\theta}|M)$$
$$\mathcal{D}|M, \boldsymbol{\theta} \sim p(\mathcal{D}|\boldsymbol{\theta}, M)$$

Marjinal olabilirlik:

$$p(\mathcal{D}|M) = \int p(\mathcal{D}|\boldsymbol{\theta}, M) p(\boldsymbol{\theta}|M) d\boldsymbol{\theta}$$

Modelin sonsal dağılımı:

$$p(M|\mathcal{D}) \propto p(M) p(\mathcal{D}|M)$$

Hedef:

$$M^* = \arg\max_M p(M|\mathcal{D}).$$

Marjinal olabilirliğin düzenleyici özelliği

Occam'ın usturası: Veriyi açıklayabilen iki modelden basit olanının marjinal olabilirliği genelde daha yüksektir.