Model - veri ilişkisi

Sinan Yıldırım

MDBF, Sabancı Üniversitesi

4 Şubat 2021

Bu sunumdaki şekiller, şu kitaplardan alınmıştır.

Pattern Recognition and Machine Learning, Christopher M. Bishop;

Mathematics for Machine Learning, Deisenroth v.d.

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

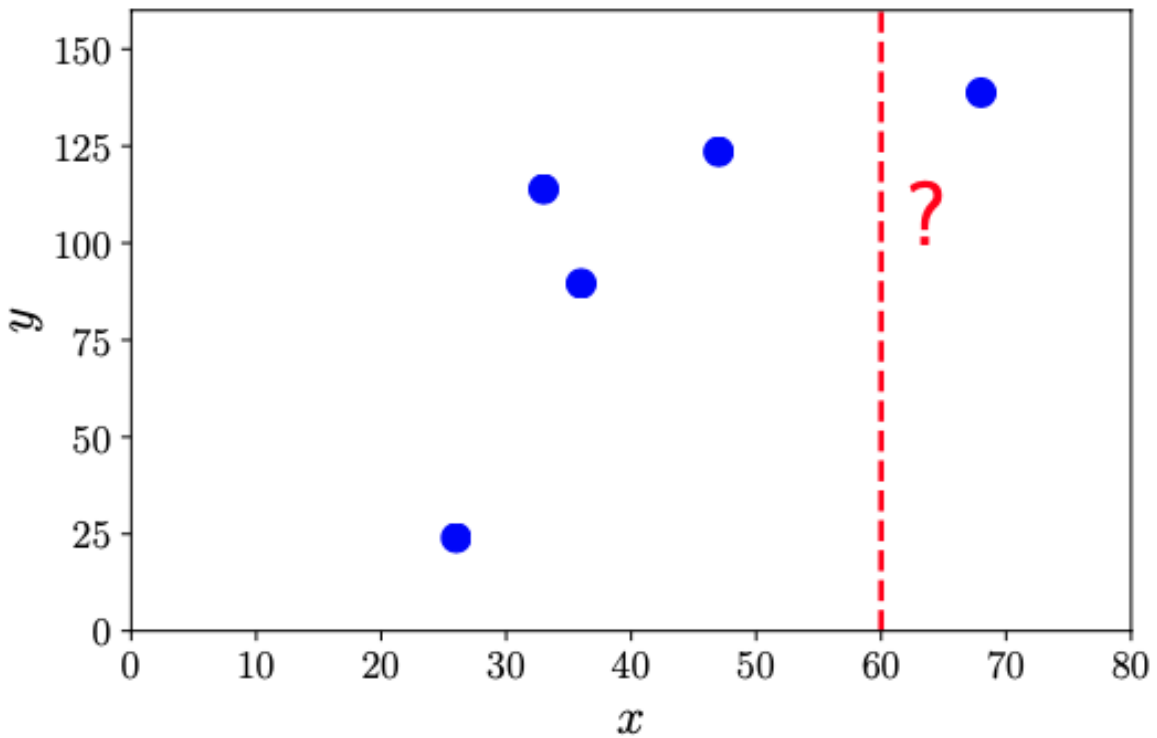
Name	Gender	Degree	Postcode	Age	Annual salary
Aditya	M	MSc	W21BG	36	89563
Bob	M	PhD	EC1A1BA	47	123543
Chloé	F	BEcon	SW1A1BH	26	23989
Daisuke	M	BSc	SE207AT	68	138769
Elisabeth	F	MBA	SE10AA	33	113888

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:

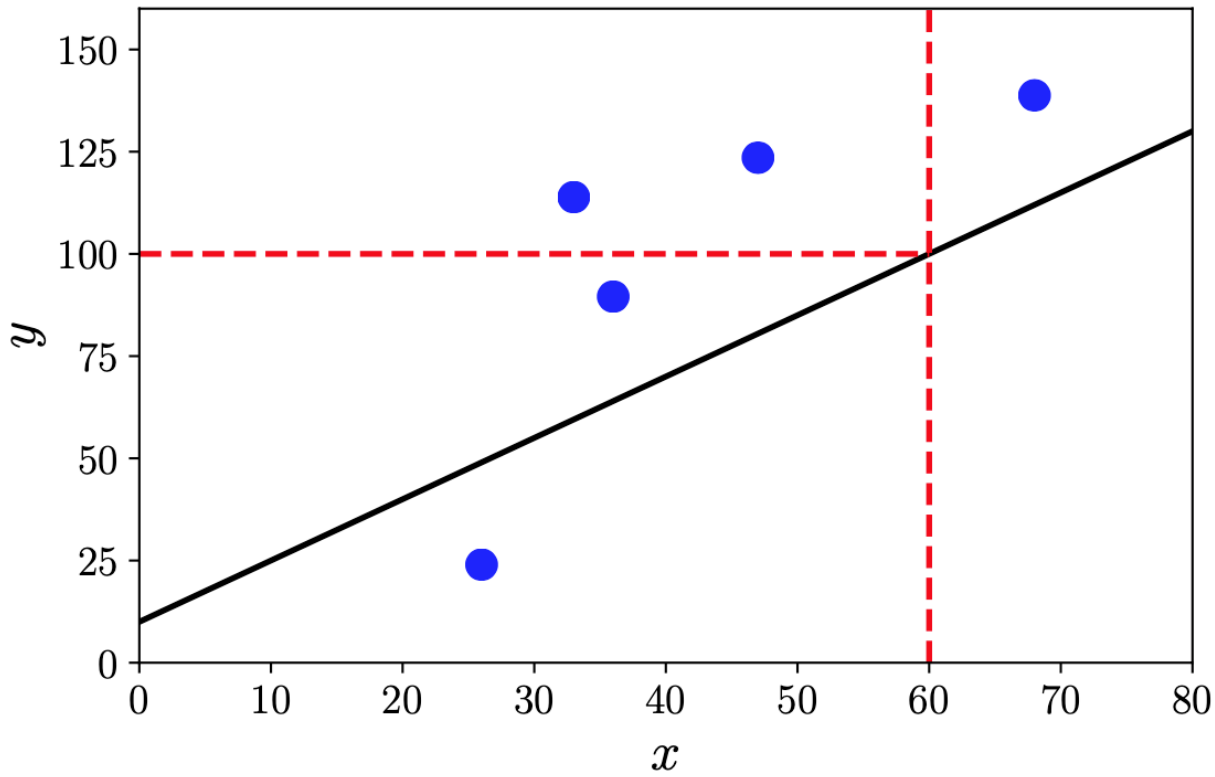
$$\ell(y_n, \hat{y}_n) = (y_n - \hat{y}_n)^2$$

Doğrusal tahminleyici:

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

Ampirik risk enküçültmesi:

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


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

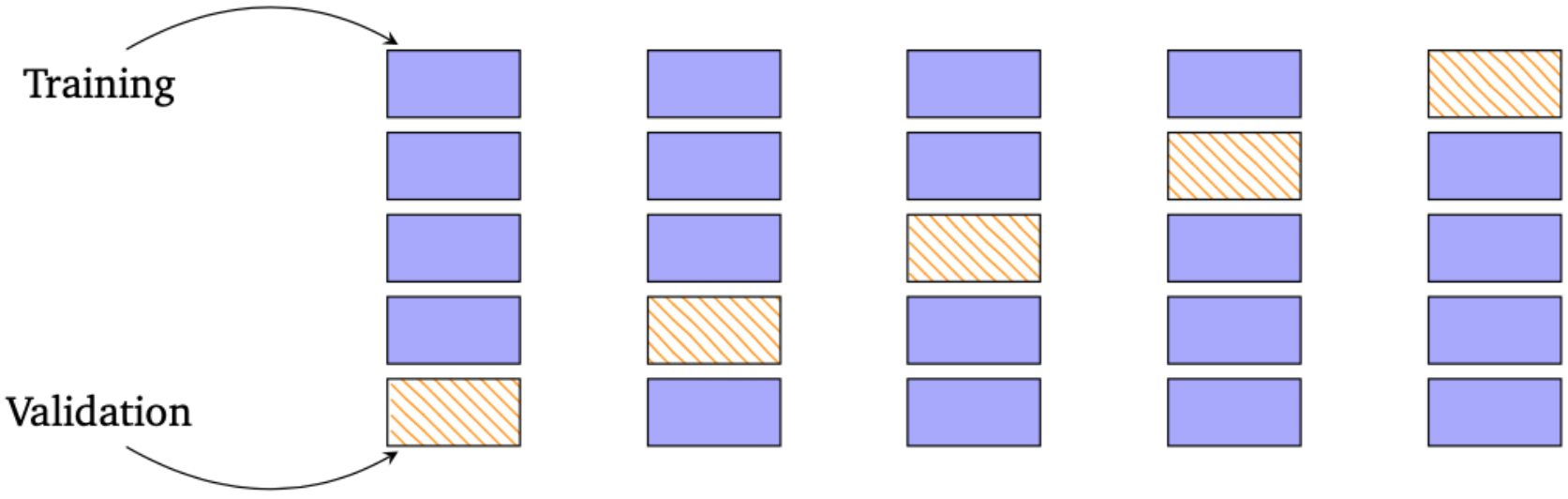
Belli bir veri üzerinde eğitilen bir model aynı kaynaktan başka verileri iyi tahminler mi?

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

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

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

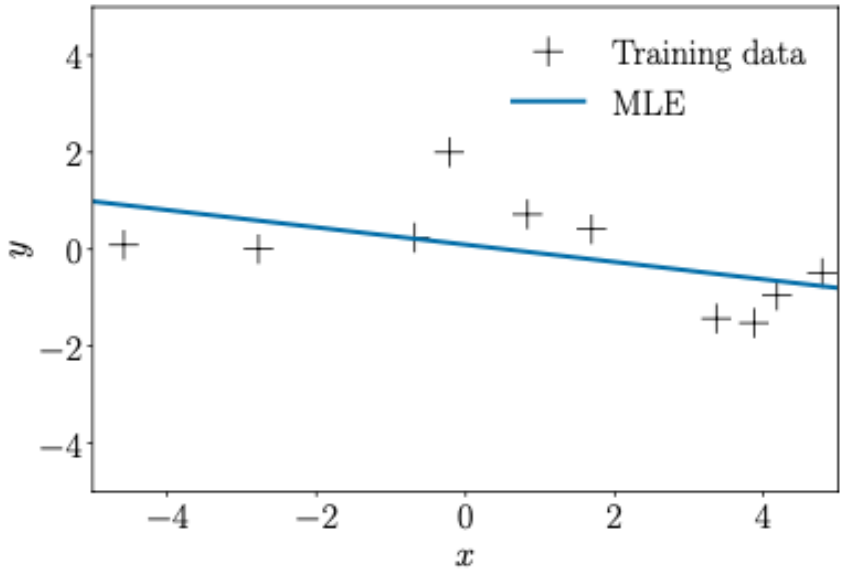
Hedef fonksiyon:

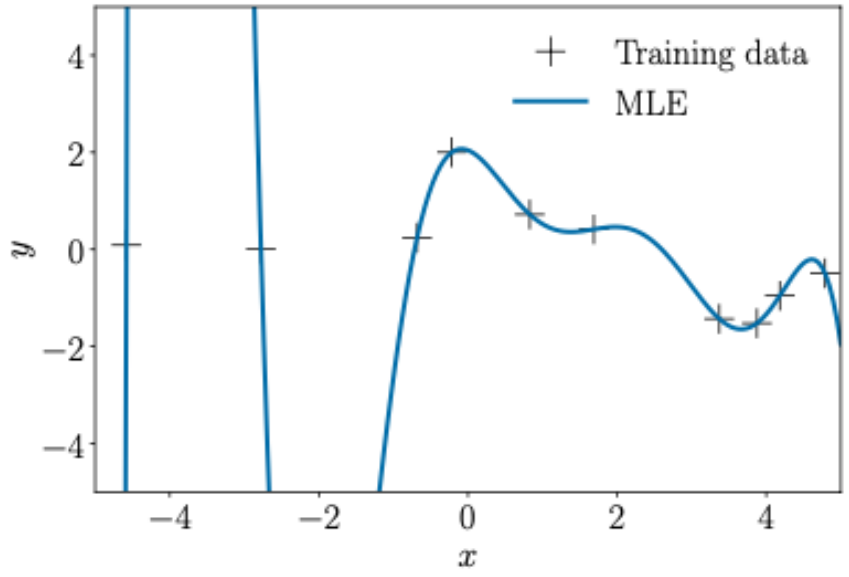
$$p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta})p(\boldsymbol{\theta})$$

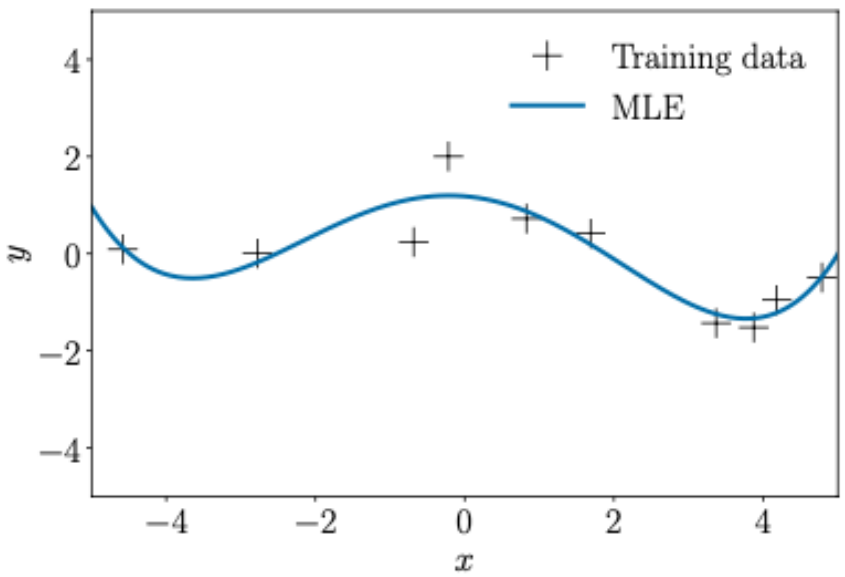
($\mathcal{X}$'in olasılık dağılımı olsa da aynı hedef kullanılabilir.)

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:

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

Noktasal kestirimler ile tahminleme:

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

Saklı değişkenli modeller

z: saklı değişkeni

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

Ortak dağılım:

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

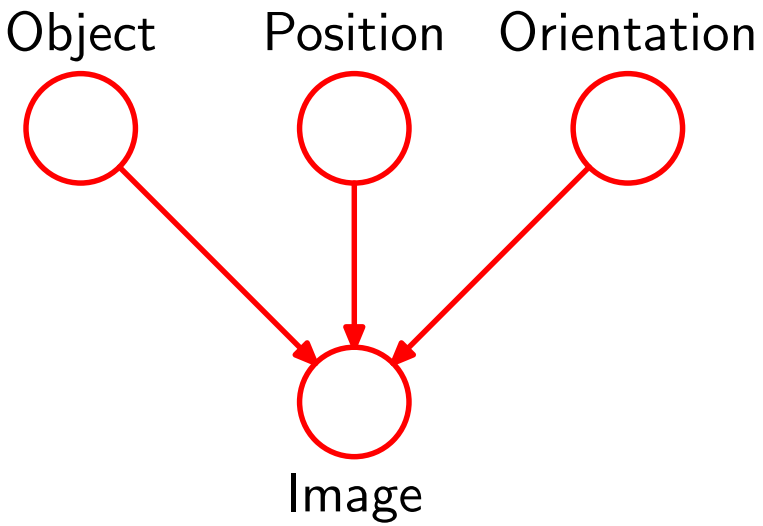
Kullanımı?

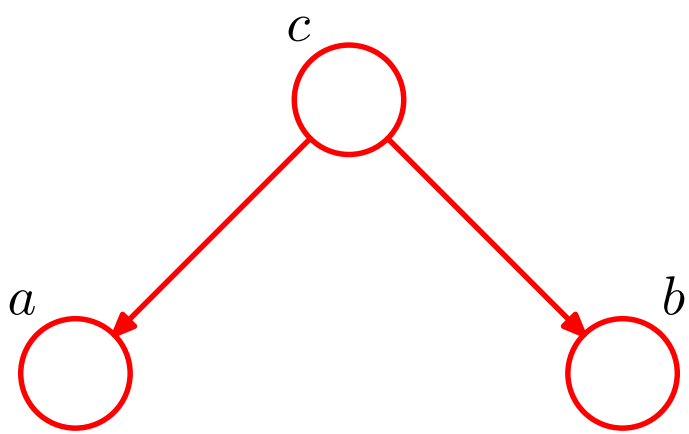
$$p(\boldsymbol{\theta}|z, \boldsymbol{x}), \quad p(z|\boldsymbol{\theta}, \boldsymbol{x})$$

Çizge modelleri

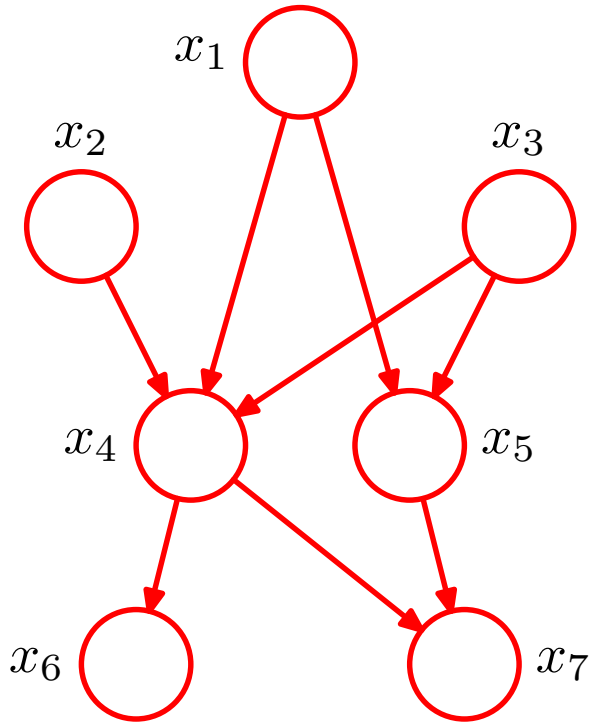
Yönlü düz çizgeler (Bayes ağları)

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

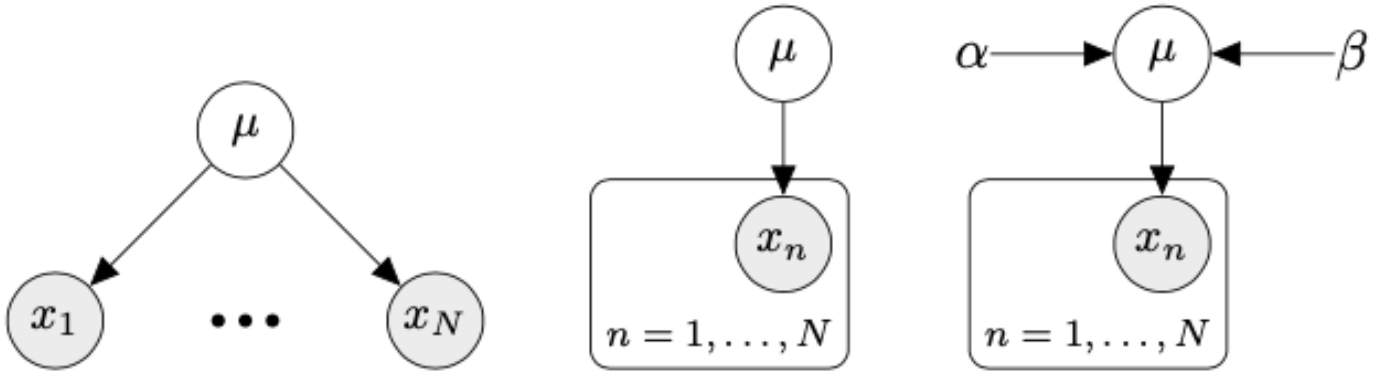
Bir madeni para N kere atılıyor. Yazı: $x = 1$, Tura: $x = 0$

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

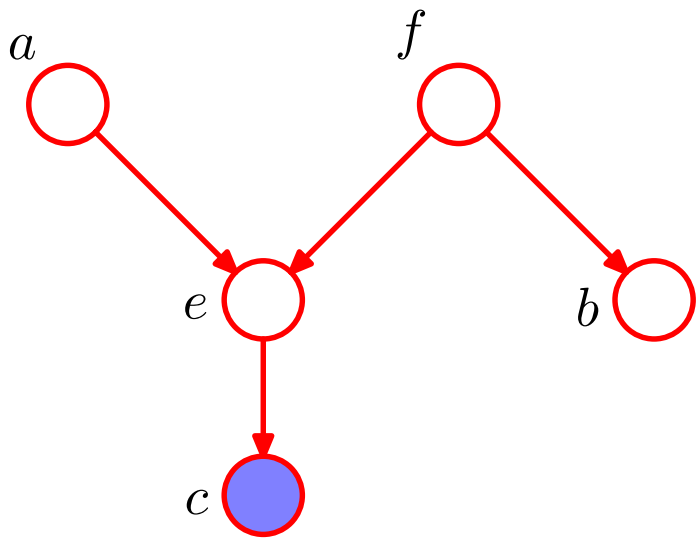
$$p(x_1, \ldots, x_N|\mu) = \prod_{n=1}^{N} p(x_n|\mu)$$

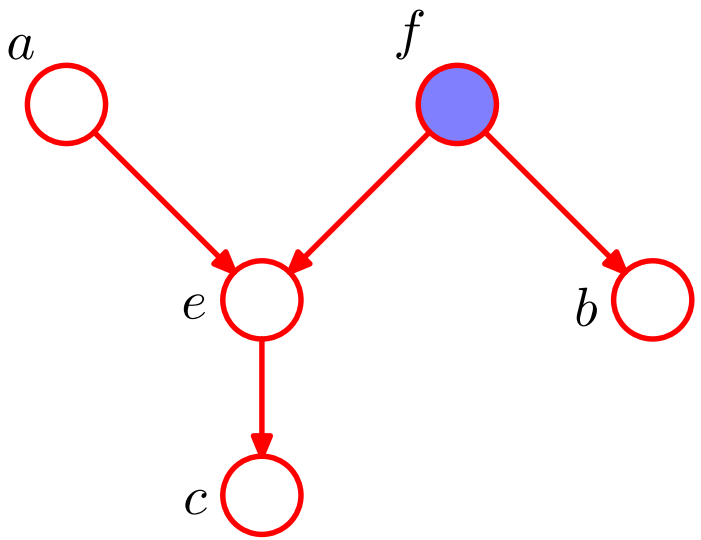
Öncül dağılım $p(\mu) = \text{Beta}(\mu|\alpha, \beta)$.

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

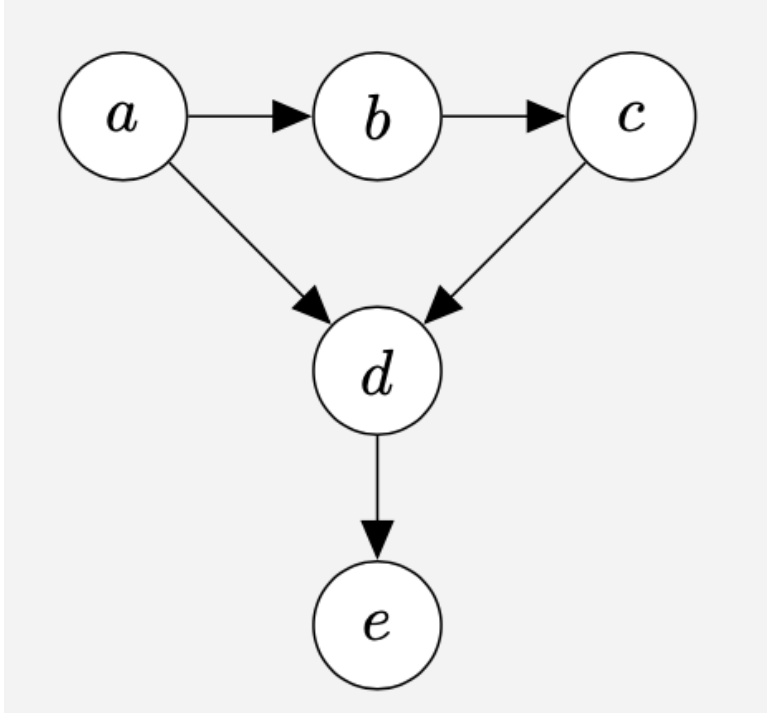
A'daki bir düğümden dan B'deki bir düğüme giden bir yol ele alalım.
Bu yolun üstündeki herhangi bir düğümü ele alalım:

- 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.



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

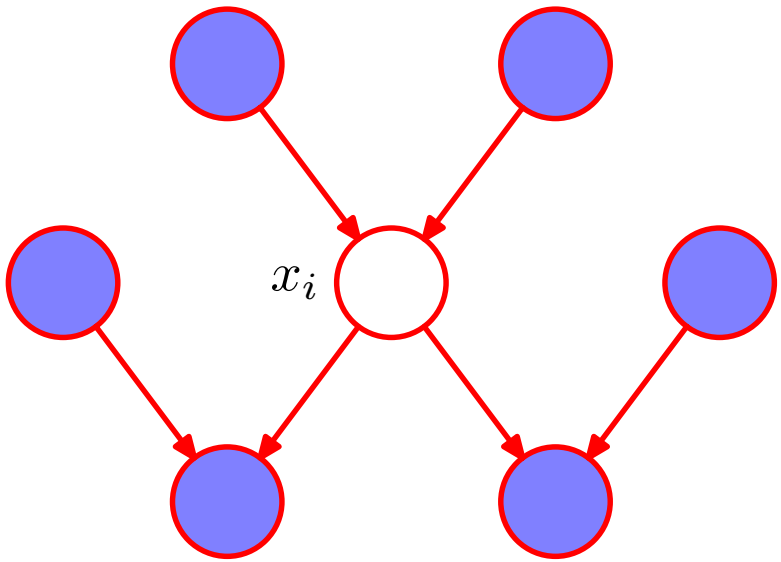
$$a \perp\!\!\!\perp c \mid a \quad ?$$

$$b \perp\!\!\!\perp d \mid c \quad ?$$

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

Markov battaniyesi

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



Bir düğüm, Markov battaniyesi verildiğinde diğer düğümlerden bağımsızdır.

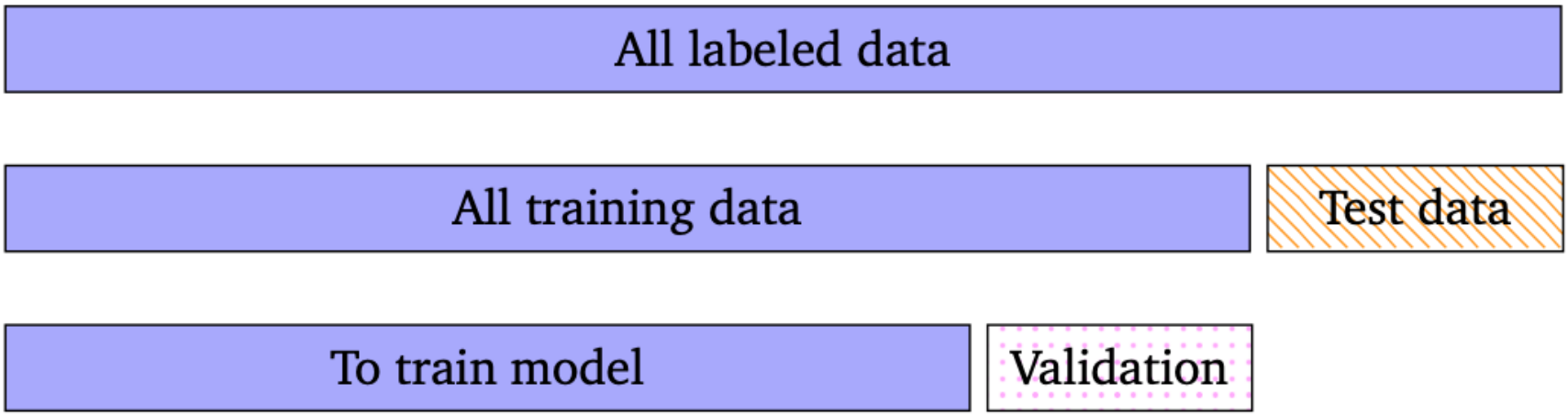
Model seçimi

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.