Subject:

Year. Month. Date.

احمدرضا فروردین — درس یادگیری ماشین

61030121 — HW 2

سوال ۱:

الف) بالا بودن بایاس نشان می‌دهد که مدل بیش از حد ساده است و داده‌ها را به خوبی مدل‌سازی نمی‌کند (under fits data) و الگوهای اساسی مدل را درک نمی‌کند.

واریانس بالا نشان می‌دهد که مدل بیش از حد پیچیده است و over fits data؛ در این حالت به نویزهای داده به جای الگوهای واقعی توجه می‌شود.

از جمله راه‌حل‌ها:

برای بایاس بالا: از مدل پیچیده‌تری استفاده شود، ویژگی‌های بیشتری اضافه کنیم یا regularization را کاهش دهیم.

برای واریانس بالا: از مدل ساده‌تری استفاده شود، ویژگی‌ها را کاهش دهیم، از cross_validation استفاده کنیم یا regularization اعمال کنیم.

توجه شود که از تکنیک‌هایی مثل Ensemble methods نیز می‌توان استفاده کرد (به عنوان مثال Bagging برای واریانس و boosting برای بایاس).

Subject:

Year. Month. Date.

ب) علت: مدل روی training data ، overfit می‌کند ، یعنی روی داده‌های تمرین خوب عمل می‌کند اما روی داده‌های test ، ضعیف عمل می‌کند.

راه حل‌ها:

* Regularization : برای محدود کردن وزن‌های مدل، جریمه‌های L_1(Lasso) یا L_2 (Ridge) را به loss function اضافه کنیم.

* cross_validation : استفاده از k-fold cross validation برای ارزیابی مدل، چرا که قابل اطمینان‌تر است.

* Feature selection : ویژگی‌های نامربوط یا ویژگی‌های تکرار شده را کاهش دهیم.

* بیشتر کردن training data برای عمومی‌سازی بهتر مدل

* از مدل ساده‌تری استفاده کنیم تا از over fitting جلوگیری شود.

پ)
$$w_{min} = \arg\min_{w} \|y - Xw\|_2^2 + \lambda \|w\|_2^2$$

$$\Rightarrow (y - Xw)^T (y - Xw) + \lambda w^T w$$

مشتق نسبت به w
$$\Rightarrow -2X^T(y - Xw) + 2\lambda w = 0$$

Subject:

Year. Month. Date.

$$\Rightarrow X^T y = (X^T X + \lambda I)\, w$$

$$\Rightarrow w_{min} = (X^T X + \lambda I)^{-1} X^T y$$

$$\begin{cases} y = X w^* + \varepsilon & \varepsilon \sim \mathcal{N}(0, \sigma^2 I) \\ w_{min} = (X^T X + \lambda I)^{-1} X^T (X w^* + \varepsilon) \end{cases}$$ (ب

$$E[w_{min}] = (X^T X + \lambda I)^{-1} X^T X w^*$$

Let $A = (X^T X + \lambda I)^{-1} X^T$:

$$w_{min} = A y = A(X w^* + \varepsilon) = \underbrace{A X w^*}_{\text{Deterministic}} + \underbrace{A \varepsilon}_{\text{Random}}$$

$$\Rightarrow Var(w_{min}) = Var[A\varepsilon] = A\, var[\varepsilon]\, A^T = \underline{\underline{A \sigma^2 A^T}}$$

بنابراین،

$$Var[w_{min}] = (X^T X + \lambda I)^{-1} X^T \sigma^2 X (X^T X + \lambda I)^{-1}$$

Subject:

Year. Month. Date.

سؤال 2 : الف)

$$\begin{cases} y_i = w^T x_i + \varepsilon_i & \varepsilon_i \underset{iid}{\sim} \mathcal{N}(0, \sigma^2) \\ P(y_i \mid x_i, w) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\dfrac{(y_i - w^T x_i)^2}{2\sigma^2}\right) \end{cases}$$

$$P(Y \mid X, w) = \prod_{i=1}^{n} P(y_i \mid x_i, w) =$$

$$(2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum (y_i - w^T x_i)^2\right)$$

$\Rightarrow$ log likelihood :

$$\log P(Y \mid X, w) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - w^T x_i)^2$$

برای بیشینه سازی log-likelihood ، منفی log-likelihood را کمینه سازی می‌کنیم، که خواهیم داشت:

$$\arg\min_{w} \sum_{i=1}^{n} (y_i - w^T x_i)^2$$

عبارت بالا برابر با SSE تابع هدف است؛ در نتیجه MLE برای رگرسیون خطی تحت توزیع گاوسی نویز، معادل با کمینه کردن SSE است.

Subject:

Year. Month. Date.

ب)

Logistic regression model:

$$P(y=1|x) = \sigma(w^T x) = \frac{1}{1+\exp(-w^T x)}$$

Binary classification : $y \in \{0,1\}$

$$P(y_i | x_i ; w) = \sigma(w^T x_i)^{y_i} (1-\sigma(w^T x_i))^{1-y_i}$$

For n i.i.d observation:

$$P(Y|X,w) = \prod_{i=1}^{n} \sigma(w^T x_i)^{y_i} (1-\sigma(w^T x_i))^{1-y_i}$$

log-likelihood maximization:

$$\log P(Y|X,w) = \sum_{i=1}^{n} [y_i \log \sigma(w^T x_i) + (1-y_i) \log(1-\sigma(w^T x_i))]$$

بیشینه سازی عبارت بالا معادل کمینه سازی منفی log-like. است :

$$\arg\min - \sum_{i=1}^{n} [y_i \log \sigma(w^T x_i) + (1-y_i) \log(1-\sigma(w^T x_i))]$$

که این تابع هزینه BCE است ؛ بنابراین MLE برای logistic regression معادل با کمینه سازی binary cross entropy است

سؤال 3 :

الف)

Feature x_1: $\{-3, -2, 3\}$ with label $+1$

Feature x_1: $\{-1, 0, 1\}$ with label -1

$\varphi(x_1) = (x_1, x_1^2)$:

class $+1$:

$(-3, +1) \longrightarrow (-3, 9)$

$(-2, +1) \longrightarrow (-2, 4)$

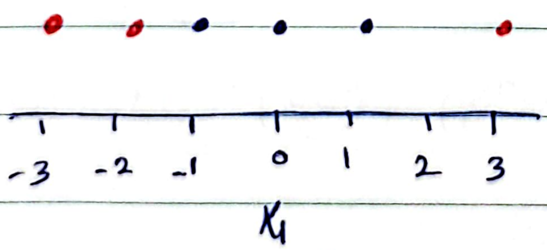
$(3, +1) \longrightarrow (3, 9)$

class -1 :

$(-1, -1) \longrightarrow (-1, 1)$

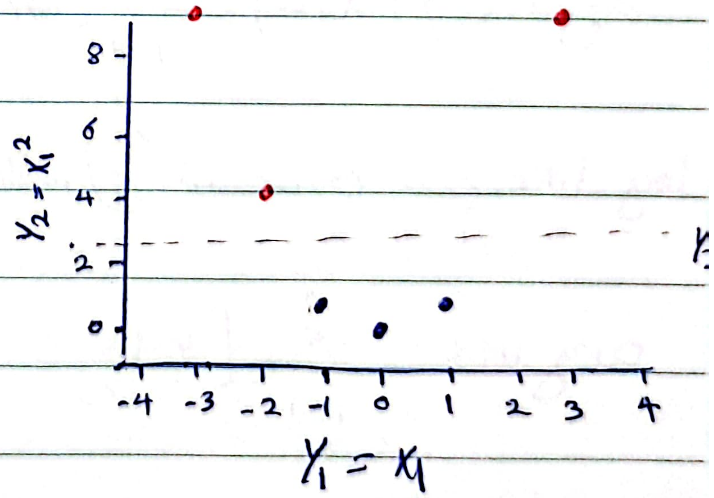
$(0, -1) \longrightarrow (0, 0)$

$(1, -1) \longrightarrow (1, 1)$



original space

به صورت خطی جدا شدنی نیست



Transformed space

به صورت خطی جدا شدنی است

class $+1$: $Y_2 \geq 4$

class -1 : $Y_2 \leq 1$

این فضا توسط خط $Y_2 = 2.5$ جدا شدنی است ولی بهینه نیست، خط بهینه در صورت بعد محاسبه شده است.

Subject:

Year. Month. Date.

ب) $K(x_1, y_1) = \varphi(x_1)\,\varphi(y_1) = x_1 y_1 + x_1^2 y_1^2$

کرنل چندجمله‌ای از درجه‌ی 2 است.

ج) support vectors:

class +1: $(-2, 4)$

class -1: $(-1, 1)$

طبق مشاهدات، optimal Hyperplane عمود منصف خط واصل بین SVs است. حال به محاسبه‌ی عمود منصف می‌پردازیم:

Midpoint = $(-1.5,\ 2.5)$

$m_{SV} = \dfrac{1-4}{-1-(-2)} = -3$

$\Rightarrow m_{Bisector} = \dfrac{1}{3}$

معادله‌ی عمود منصف:

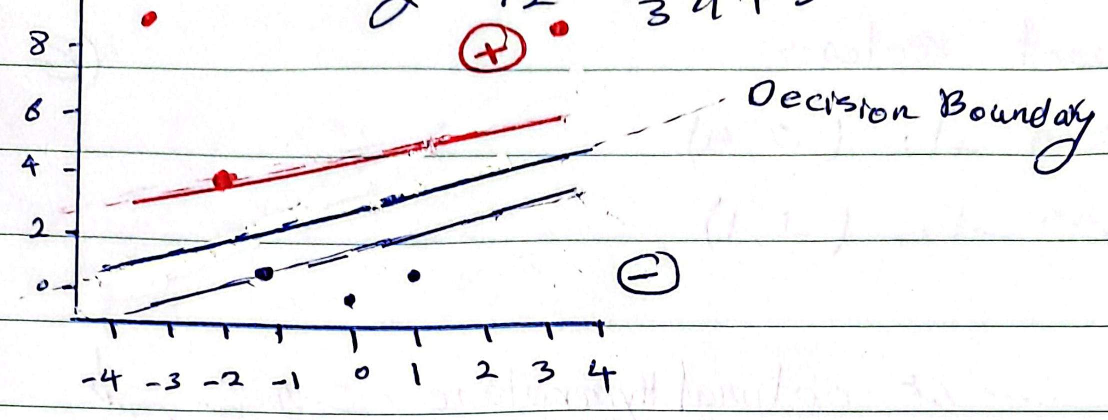
$y_2 - 2.5 = \frac{1}{3}\left(y_1 - (-1.5)\right) \Rightarrow y_2 = \frac{1}{3} y_1 + 3$

Margin = $\sqrt{(-1)^2 + (3)^2} = \sqrt{10} \simeq 3.16$

Plus plane: $Y_2 = \frac{1}{3} y_1 + \frac{2}{3} + 4$

Minus plane: $Y_2 = \frac{1}{3} Y_1 + \frac{1}{3} + 1$

Decision Boundary: $Y_2 = \frac{1}{3} Y_1 + 3$



Decision Boundary in Original space:

$Y_2 = \frac{1}{3} Y_1 + 3$

$\Longrightarrow$

$$X_1^2 = \frac{1}{3} X_1 + 3 \Rightarrow X_1^2 - \frac{1}{3} X_1 - 3 = 0$$

$$\Rightarrow X_1 = \frac{\frac{1}{3} \pm \sqrt{(\frac{1}{3})^2 + 12}}{2} = -1,54 \; , \; 1,95$$

Class +1 $\Leftarrow$ $X_1 > 1,95$ یا $X_1 < -1.54$ [illegible]

else class -1

Veri: Freation:

Class +1 Points: -3, -2, 3 (all outside $[-1,54, 1.95]$)

Class -1 Points: -1, 0, 1 (all inside $[-1,54, 1,95]$)

Subject:

Year. Month. Date.

هـ : برای رسیدن به فرم کلی مسئله‌ی دوگان برای hard svm باید Dual کردن مسئله از آنجا آغاز می‌شود که مسئله‌ی بهینه سازی اولیه‌ای داشتیم (maximizing margin) و با استفاده از Lagrangian duality

آن را به فرم جدیدی در آوریم تا حل کردن آن راحت‌تر باشد:

① Primal Problem:

هدف پیدا کردن hyperplane است که $w\varphi(x)+b=0$ که فاصله‌ی بین کلاس‌ها را بیشینه می‌کند.

قید و محدودیت ما این است که همه‌ی نقاط در شرط زیر صدق کنند:

$$y_i\left(w\varphi(x_i)+b\right) \geq 1$$

② Lagrangian Duality:

برای هر قید ضرایب لاگرانژ $\alpha_i \geq 0$ را تعریف می‌کنیم.

فرمول لاگرانژ، تابع هدف و قیود را ترکیب می‌کند:

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum \alpha_i\left[y_i\left(w\varphi(x_i)+b\right)-1\right]$$

$$\begin{cases} \text{minimize } L \text{ w.r.t } w, b \\ \text{maximize w.r.t } \alpha_i \end{cases}$$

③ Dual Problem (Final Form):

بعد از حذف w, b مسئله‌ی دوگان برابر خواهد بود با:

Subject:

Year. Month. Date.

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j \underbrace{K(x_i, x_j)}_{\phi(x_i)\,\phi(x_j)}$$

Kernel

$$s.t: \alpha_i \geq 0 \;,\; \sum_{i=1}^{n} \alpha_i y_i = 0$$

- چرا SVs مهم هستند؟

SVs نقاطی با $\alpha_i > 0$ هستند، آنها در margin قرار گرفتند؛ یعنی:

$$y_i(w\,\phi(x_i) + b) = 1$$

باید توجه کرد که تنها SVs در decision boundary نقشی دارند چرا که نقاط غیر SV، $\alpha_i = 0$ دارند و در $\sum_i \alpha_i y_i \phi(x_i)$ تأثیری ندارند.

پیش بینی که با کرنل بدست می‌آید:

$$y(x) = \text{sign}\Big(\sum_{i \in SV} \alpha_i y_i K(x_i, x) + b\Big)$$

تنها به SVs بستگی دارد و از آنها استفاده می‌کند.

در فرمول دو ترم داریم:

$\sum \alpha_i$: بیشینه کردن این ترم به معنای تشویق کردن مدل به داشتن نقاط SV بیشتر است (با شرط قیود محدود می‌شود)

$\frac{1}{2} \sum \alpha_i \alpha_j y_i y_j K(x_i, x_j)$: مقادیر بزرگ α را برای نقاط مشابه جریمه می‌کند؛ در واقع بیشینه سازی margin را حمایت می‌کند.

Subject:

Year. Month. Date.

Constraints :

اجرای شرایط KKT ($\alpha_i \geq 0$) و $\underline{\sum \alpha_i y_i = 0}$ -

ایجاد تعادل بین کلاس ها

ج)

$$\begin{cases} \text{SVs (in original space):} \\ x_1 = -2 \text{ (class } y_1 = +1) \\ x_2 = -1 \text{ (class } y_2 = -1) \\ K(x_1, y_1) = x_1 y_1 + x_1^2 y_1^2 \end{cases}$$

مقادیر kernel را حساب می‌کنیم:

$K(-2, -2) = 20$

$K(-2, -1) = 6$

$K(-1, -1) = 2$

مسئله‌ی دوگان برابر خواهد بود با:

$$\max W(\alpha) = \alpha_1 + \alpha_2 - \frac{1}{2}\left(20\alpha_1^2 - 12\alpha_1\alpha_2 + 2\alpha_2^2\right)$$

constraints: $\begin{cases} \alpha_1 - \alpha_2 = 0 \\ \alpha_1, \alpha_2 \geq 0 \end{cases}$

حل این مسئله:

$$\alpha_1 = \alpha_2 = \alpha \Rightarrow W(\alpha) = 2\alpha - \frac{1}{2}\left(20\alpha^2 - 12\alpha^2 + 2\alpha^2\right) = 2\alpha - 5\alpha^2$$

$$\frac{dW}{d\alpha} = 2 - 10\alpha = 0 \Rightarrow \alpha = 0.2 \Rightarrow \boxed{\alpha_1 = \alpha_2 = 0.2}$$

Subject:
Year. Month. Date.

ز)

محاسبه‌ی بایاس b :

Using SV $(-2, +1)$:

$$y_1 \left(\sum_{j=1}^{2} \alpha_j y_j K(u_1, u_j) + b \right) = 1$$

$$\xrightarrow{\text{جای‌گذاری}} +1\left(0.2(+1)20 + 0.2(-1)6 + b\right) = 1$$

$$\Rightarrow b = -1.8$$

حال برای تایید این بایاس از SV $(-1,-1)$ استفاده می‌کنیم :

$$-1\left(0.2(+1)6 + 0.2(-1)2 + (-1.8)\right) = -1(-1) = 1$$

(valid)

حال به محاسبه‌ی Decision Boundary در فضای اصلی می‌پردازیم :

Classifier :

$$y(x) = \text{sign}\left(0.2(+1)K(x,-2) + 0.2(-1)K(x,-1) - 1.8\right)$$

برای بدست آوردن مرز : $y(x) = 0$:

$$0.2K(x,-2) - 0.2K(x,-1) - 1.8 = 0$$

Subject:

Year. Month. Date.

جای گذاری کردن:

$$0.2(-2x+4x^2)-0.2(-x+x^2)-1.8=0$$

$$\Rightarrow 0.6x^2-0.2x-1.8=0$$

$$\Rightarrow x=\frac{0.2\pm\sqrt{0.04+4.32}}{1.2}\approx \underbrace{-1.56,\ 1.89}_{*}$$

از طرفی مرز هندسی در فضای تبدیل ($Y_2=\frac{1}{3}Y_1+3$) معادل است با:

$$Y_1^2=\frac{1}{3}Y_1+3 \Rightarrow Y_1^2-\frac{1}{3}Y_1-3=0$$

$$\Rightarrow Y_1\approx \underbrace{-1.54,\ 1.95}_{**}$$

از آنجا که * و ** با تقریب خوبی، با هم برابر هستند لذا تأییدی بر معادل بودن مرز هندسی و مرز بیزی است که بدست آوردیم.

سوال 4 :

الف) در هر دو روش hard SVM و soft SVM ، تابع هدف برای بهینه سازی یکی است، تفاوت در Constraints آنهاست.
در روش soft-SVM ، کران بالای C را برای α_i در نظر می گیریم. این کار باعث می شود مدل در شرایطی که عدم قطعیت وجود دارد انعطاف پذیری به خرج دهد (allows slack).
در واقع در hard margin SVM ، فرض بر این است که داده ها به طور کامل جدایی پذیر هستند در حالیکه روش Soft_Margin داده های misclassified را با متغیر های slack (ε_i) کنترل می کند.

ب)

x_1 : $\alpha_i = 0$

x_2 : $0<\alpha_i<C$

x_3 : $\alpha_i = C$

x_4 : $\alpha_i = C$

ج) خیر، از جمله دلایل می توان گفت:

۱. بعد های نامتناهی: تبدیل $\phi(x)_\infty$ ویژگی های نامتناهی تولید می کند
$$\left(e^{-\frac{x^2}{2}} \frac{x^i}{\sqrt{i!}}\right)_{i=0}^{\infty}$$

۲. محدودیت های محاسباتی: کامپیوتر های امروزی با حافظه ی محدود، توان

Subject:

Year. Month. Date.

ذخیره سازی ترم های نامحدود را ندارند.
از طرفی الگوریتم ها امکان پردازش بردارهای با بعدهای نامتناهی را به طور مستقیم ندارند.

ج)

$$K(x,y) = \phi_\infty(x)^T \phi_\infty(y).$$

$$= \sum_{i=0}^{\infty} \left(\frac{e^{-\frac{x^2}{2}} x^i}{\sqrt{i!}} \right) \left(\frac{e^{-\frac{y^2}{2}} y^i}{\sqrt{i!}} \right)$$

$$= e^{-(x^2+y^2)/2} \sum_{i=0}^{\infty} \frac{(xy)^i}{i!}$$

$$\Longrightarrow \sum_{i=0}^{\infty} \frac{(xy)^i}{i!} = e^{xy}$$

$$\Longrightarrow K(x,y) = e^{-\frac{x^2+y^2-2xy}{2}} = e^{-\frac{(x-y)^2}{2}}$$

$$\Longrightarrow K(x,y) = e^{-(x-y)^2/2}$$

$K(x,y)$ بدست آمده Kernel Gaussian(RBF) با $\sigma=1$ است
این Kernel امکان جداسازی خطی در R^∞ را بدون اعمال مستقیم تبدیل می‌دهد.

Ehsan

Subject:

Year. Month. Date.

سؤال ۵:

(قسمت الف)

1. Initial Entropy :

$$Entropy(S) = -\sum p_i \log_2 p_i = -\left(\frac{4}{7}\log_2\frac{4}{7} + \frac{3}{7}\log_2\frac{3}{7}\right)$$

$\begin{cases} 4 \text{ Yes} \\ 3 \text{ No} \end{cases}$ ≈ 0.985 bits

2. Information Gain (IG) for Attributes:

a) outlook :

sunny : 2 No → Entropy = 0

overcast : 2 yes → Entropy = 0

Rain : 2 Yes , 1 No → $Entropy = -\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right)$ ≈ 0.918

$\Rightarrow IG(S, outlook) =$

$.985 - \left(\frac{2}{7}\times 0 + \frac{2}{7}\times 0 + \frac{3}{7}\times .918\right) \approx$

$.246$

b) Temperature :

Hot : 1 Yes , 2 No → Entropy = .918

Mild : 2 Yes , 1 No → Entropy = .918

Cool : 1 Yes → Entropy = 0

$\Rightarrow IG(S, Temperature) = .985 - \left(\frac{3}{7} .918 + \right.$

Ehsan

Subject:

Year. Month. Date.

$\frac{3}{7}$,918 + $\frac{1}{7}$ x 0) ≃ ,020

c) wind:

weak: 3 Yes, 1 No → Entropy ≈ ,811

strong: 1 Yes, 2 No → Entropy ≈ ,918

⇒ IG (S, wind) =

,985 − ($\frac{4}{7}$ x ,811 + $\frac{3}{7}$,918) ≈ ,020

⟶ Root Node selection:

outlook چون با ،246 IGs بالاترین IG را داره

تقسیم به زیرشاخه‌ها:

Subset: Rain: 2 Yes, 1 No

Entrop Rain ≈ ,918

IG Calculations for Rain:

☑ Temperature:

* cool: 1 Yes, 1 No → Entropy = 1

* Mild: 1 Yes → Entropy = 0

⇒ IG (Rain, Temperature) =

,918 − ($\frac{2}{3}$ x1 + $\frac{1}{3}$ x0) ≈ ,252

☑ wind:

Ehsan

Subject:

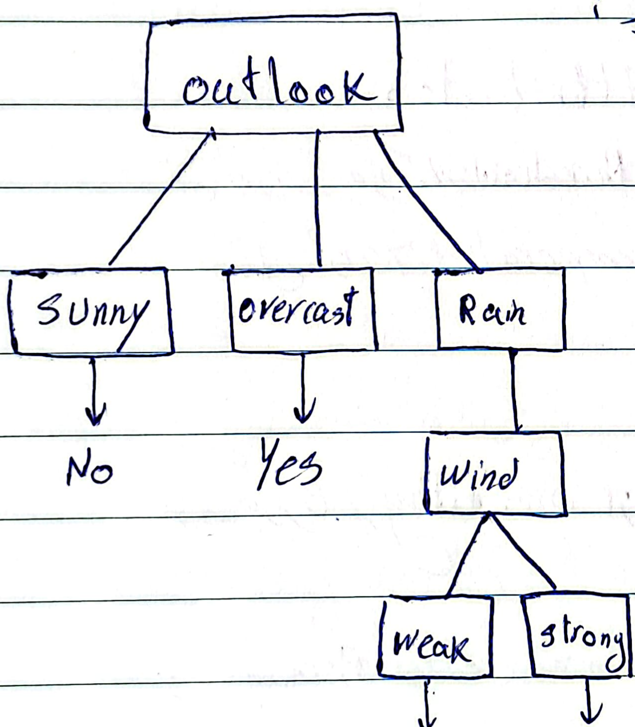
Year. Month. Date.

* Weak: 2 Yes → Entropy = 0
* strong: 1 No → Entropy = 0

$\Rightarrow IG(Rain, Wind) = 0.918 - (\frac{2}{3} \times 0 + \frac{1}{3} \times 0) = 0.918$

بنابراین $IG(Rain, Wind)$ از $IG(Rain, Temp.)$ بیشتر است لذا تقسیم بندی برای شاخه ی Rain روی Wind خواهد بود:

درخت تصمیم خواهد بود:



دقت درخت تصمیم روی Training Data: $\frac{7}{7} = 100\%$

دقت درخت تصمیم روی Test Data: $\frac{3}{4} = 75\%$

Subject:

Year. Month. Date.

ب)

گام های الگوریتم C4.5 :

1. Gain Ratio Calculation:

$$\text{GainRatio}(A) = \frac{IG(A)}{\text{SplitInfo}(A)} \quad , \quad \text{SplitInfo}(A) = -\sum \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

جهت جلوگیری از سوگیری به سمت ویژگی های با کاردینالیتی بالا ، IG را نرمال سازی می کند.

2. Handling Continuous Attributes:

به صورت خودکار optimal thresholds را انتخاب می کند، برای مثال Temperature > 25°

3. Pruning :

برای جلوگیری از overfitting از PEP استفاده می کند. PEP یا Pessimistic Error Pruning روش و تکنیک از نوع Post_Pruning است که با حذف شاخه هایی از درخت تصمیم که الگوهایی را از داده های آموزشی یاد گرفته اند که عمومیت ندارند یا آنکه از داده های noise یادگیری خود را انجام داده اند، به کاهش overfitting منجر می شود.

Subject:

Year. Month. Date.

4. Missing Values:

نمونه ها را به طور متناسب در بین شاخه ها توزیع می کند.

از مزایای C4.5 نسبت به ID3 می توان به موارد زیر اشاره کرد:

1. Numeric Attribute support:

ID3 توانایی کنترل کردن داده های پیوسته را ندارد، این در حالی است که C4.5 داده های عددی را بطور پویا تفکیک می کند.

2. Robust Splitting Criterion (Gain Ratio):

روش ID3 ویژگی های با مقادیر بالا را ترجیح می دهد در حالیکه در روش C4.5، GainRatio چنین ویژگی هایی را با SplitInfo جریمه می کند.

3. Missing Data Tolerance:

C4.5 نمونه هایی که فاقد برخی از ویژگی ها باشند را از دست نمی دهد.

4. overfitting Mitigation:

C4.5 شامل Post_Pruning می شود این در حالی است که ID3 درخت تقسیم را به طور کامل می سازد.