به نام خدا

یادگیری ماشین
نیم‌سال دوم ۱۴۰۳-۱۴۰۴
دکتر ربیعی و دهقانیان

دانشکده مهندسی کامپیوتر

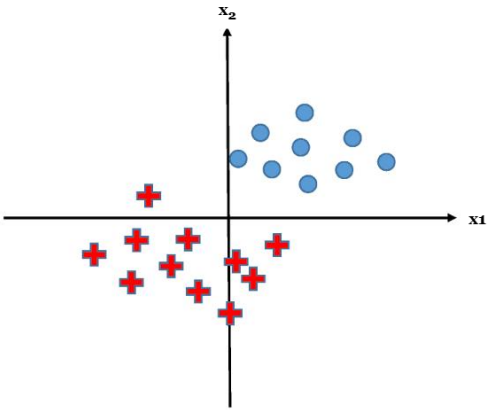
آزمون پایانترم

زمان امتحان: ۱۸۰ دقیقه

بخش مفهومی (۲۵ نمره)

سوال ۱) برای هر یک از این سوالات این بخش، در یک پاراگراف توضیحات خواسته شده را ارائه دهید. نمره‌ی تمام سوالات این بخش با یکدیگر برابر است.

- [۱.۱] ما با مسئله‌ای روبه‌رو هستیم که هدف آن، تخمین امتیاز یک کامنت درباره‌ی غذای یک رستوران بر اساس ویژگی‌های استخراج‌شده از آن کامنت است. این امتیاز یکی از اعداد صحیح بین ۱ تا ۵ است. سوالی که مطرح می‌شود این است که آیا این مسئله، یک مسئله‌ی دسته‌بندی [۱] است یا رگرسیون ؟
- [۲.۱] به صورت مختصر توضیح دهید که تفاوت داده‌ی تست و ارزیابی [۲] چیست؟ چرا از داده‌های تست برای ارزیابی استفاده نمی‌کنیم؟
- [۳.۱] مجموعه داده‌ی آموزشی $S = \{(x_i, y_i)\}_{i=1}^{m}$ داده شده است که x_i و y_i ها حقیقی هستند. کدام‌یک از دو مدل $h_1(x) = w_1^2 x + w_2 x$ و $h_2(x) = wx$ روی مجموعه‌ی آموزش، خطای تجربی کمتری را خواهد داشت؟ توضیح دهید.
- [۴.۱] مسئله‌ی یادگیری‌ای را در نظر بگیرید که در آن هر داده‌ی نمونه، دارای n ویژگی به صورت $x_1, x_2, \ldots, x_n$ است که هر یک از x_i ها فقط می‌تواند مقدار ۰ یا ۱ داشته باشد.مدلی بر پایه‌ی پرسپترون طراحی کنید که اگر تعداد ویژگی‌هایی که مقدارشان ۰ است بیشتر از تعداد ویژگی‌هایی با مقدار ۱ باشد، خروجی مدل برابر ۱ باشد؛ و در غیر این صورت، خروجی مدل ۰ باشد.
- [۵.۱] فرض کنید تابع رگرسیون لجستیک [۳] به صورت $h(x) = \frac{1}{1+\exp(-w_1 x_1 - w_2 x_2)}$ برای دسته‌بندی داده‌های زیر استفاده کنیم. فرض کنید که به **تابع هزینه‌ی** رگرسیون لجستیک عبارت $c(|w_1|+|w_2|)$ اضافه گردد. اگر پارامتر c به تدریج از مقادیر نزدیک به صفر به سمت بی‌نهایت افزایش یابد، کدام‌یک از اتفاقات زیر رخ می‌دهد؟ علت انتخاب خود را توضیح دهید. سمت صفر میل نمی‌کنند.



- [۱-] هر دوی w_1, w_2 هم‌زمان به صفر میل می‌کنند.
- [۲-] نخست w_1 صفر می‌شود و سپس w_2 به سمت صفر میل می‌کند.
- [۳-] نخست w_2 صفر می‌شود و سپس w_1 به سمت صفر میل می‌کند.
- [۴-] هیچ کدام از w_1, w_2 مقادیر با افزایش c به

بخش تئوری (۷۵ نمره)

سوال ۲) مجموعه دادگان زیر را در نظر بگیرید: با استفاده از این دادگان قصد داریم که ویژگی species را برای داده‌ی زیر

[1] classification
[2] validation
[3] logistic regression

Species	Smelly	Height	Legs	Color	No. Sl.
M	Yes	Short	۳	White	۱
M	No	Tall	۲	Green	۲
M	Yes	Short	۳	Green	۳
M	Yes	Short	۳	White	۴
H	No	Short	۲	Green	۵
H	No	Tall	۲	White	۶
H	No	Tall	۲	White	۷
H	Yes	Short	۲	White	۸

مشخص نماییم:

$$X = \{\text{Color} = \text{Green}, \text{Legs} = 2, \text{Height} = \text{Tall}, \text{Smelly} = \text{No}\}$$

با توجه به دادگان فوق، رویکرد مناسب برای حل این مسئله چیست؟ علت انتخاب این رویکرد را توضیح دهید. سپس با استفاده از رویکردی که انتخاب نموده‌اید، مسئله را حل نمایید.

سوال ۳) طول عمر یک ماشین را می‌توان به شکل یک متغیر تصادفی X از توزیع نمایی با پارامتر مجهول θ مدل کرد، به‌طوری که $p(x|\theta) = \theta e^{-\theta x}$ برای $x \geq 0$ و $\theta > 0$ برقرار باشد.

- برای مجموعه دادگان $\mathcal{D} = \{x_i\}_{i=1}^{N}$ که فرض i.i.d بودن برای آن‌ها برقرار است. برآوردگر بیشینه درست‌نمایی [۴] را برای پارامتر θ به‌دست آورید.
- فرض کنید که می‌دانیم پارامتر θ نیز از توزیع پیشین [۵] نمایی به شکل $p(\theta|\lambda)$ به دست می‌آید که λ پارامتر ثابت و داده‌شده است. توزیع پسین [۶] $p(\theta|\mathcal{D}, \lambda)$ را به دست آورید.

سوال ۴) در ماتریس $X \in R^{n \times d}$ هر سطر مربوط به یک داده‌ی یادگیری می‌باشد (که هر کدام دارای d ویژگی هستند) و بردار $\mathbf{y} \in R^n$ بردار خروجی‌های مورد نظر و بردار $\mathbf{w} \in R^d$ ضرایب رگرسیون می‌باشند که $\mathbf{w}^*$ پارامترهای بهینه برای ماتریس ویژگی‌های X می‌باشد. در این‌جا برای سادگی در محاسبات فرض می‌کنیم که داده‌ها سفید شده‌اند بدین معنی که $X^T X = I$ می‌باشد. در رگرسیون LASSO فرض بر این است که ضرایب بهینه از رابطه‌ی زیر به‌دست می‌آیند:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} J_\lambda(\mathbf{w}),$$

$$J_\lambda(\mathbf{w}) = \frac{1}{2}\|\mathbf{y} - X\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1 \quad (\lambda > 0)$$

- [۱.] ابتدا نشان دهید که که سفید کردن داده‌های آموزش باعث مستقل شدن ویژگی‌های داده‌های آموزشی می‌شود بدین معنا که می‌توان به‌طور مستقل w_i^* را تنها از ویژگی i ام و خروجی تعیین نمود. برای نشان دادن این مطلب، نشان دهید که می‌توان $J_\lambda(\mathbf{w})$ را به صورت زیر نوشت:

$$J_\lambda(\mathbf{w}) = g(\mathbf{y}) + \sum_{i=1}^{d} f(X_{.,i}, \mathbf{y}, w_i; \lambda),$$

که در آن $X_{.,i}$ نشان‌دهنده ستون iام ماتریس X است و $\mathbf{y}$ ماتریس پاسخ و f تابعی است که به $X_{.,i}$، y، w_i، λ وابسته است.

[4] maximum likelihood estimator
[5] prior
[6] posterior

- [۲.] اگر $w_i^* > 0$ ، مقدار w_i^* را به‌دست آورید.
- [۳.] اگر $w_i^* < 0$ ، مقدار w_i^* را به‌دست آورید.
- [۴.] با توجه به دو قسمت قبل، در چه شرایطی مقدار $w_i^* = 0$ می‌شود؟ چگونه می‌توان این شرط را اعمال نمود؟

سوال ۵) فرض کنید در یک مساله پیشبینی قیمت خانه در شهر تهران بر اساس متراژ، از مدل رگرسیون خطی استفاده کرده ایم. x متراژ خانه بر حسب متر مربع، y مقدار واقعی قیمت خانه و $\hat{y}$ مقدار پیشبینی شده توسط مدل است. فرمول خطای تعمیم پذیری [۷] به صورت $E[(y-\hat{y})^2]$ به شما داده شده است.

- [۱.] نشان دهید این خطا را می توان به سه بخش اریبی [۸]، واریانس و نویز تجزیه کرد و بر اساس آن مفهوم تعادل اریبی واریانس [۹] را توضیح دهید . فرض کنید $y = f(x) + \epsilon$ که ϵ نویز با $E(\epsilon) = 0$ و $var(\epsilon) = \sigma^2$ است.
- [۲.] اگر در مدل رگرسیون خطی از منظم‌سازیِ L_2 [۱۰] استفاده کنیم، توضیح دهید این عمل چه تاثیری روی تعادل اریبی و واریانس می‌گذارد و چگونه ممکن است به بهبود تعمیم پذیری مدل کمک کند؟
- [۳.] فرض کنید به جای استفاده از تمامی داده های موجود، مدل رگرسیون خطی را تنها روی خانه های منطقه ۱ تهران (زیر مجموعه ای از مجموعه داده) آموزش داده ایم. توضیح دهید چگونه آموزش مدل در این حالت بر مقدار اریبی و واریانس مدل تاثیر می‌گذارد؟ بر اساس بخش اول توضیح دهید چرا خطای تعمیم پذیری در این آزمایش ممکن است افزایش یابد؟
- [۴.] اگر به جای مدل خطی، برای پیشبینی قیمت از یک مدل چند جمله ای با درجه بالاتر از ۱ استفاده کنیم، با ذکر دلیل بیان کنید افزایش درجه چند جمله ای چه تاثیری بر میزان اریبی و واریانس دارد؟

[7] generalization error
[8] bias
[9] bias variance tradeoff
[10] L_2 regularization

پاسخ‌نامه

سوال ۱

● [۱.۱ (۵ نمره)] در این‌جا با یک مسئله‌ی **رگرسیون** مواجه هستیم. دلیل اصلی این است که مقادیر ممکن برای خروجی (یعنی امتیازهای ۱ تا ۵) نه تنها گسسته هستند، بلکه دارای *ترتیب* و *معنای* عددی نیز می‌باشند. به‌عبارت دیگر، فاصله‌ی بین امتیاز ۲ و ۴ بیشتر از فاصله‌ی بین امتیاز ۲ و ۳ است و این خود نشان‌دهنده‌ی وجود یک ساختار ترتیبی و عددی در خروجی‌هاست.

در مقابل، در مسائل *دسته‌بندی*[11] کلاس‌ها صرفاً برچسب‌هایی بدون ترتیب ذاتی هستند؛ مثلاً در مسئله‌ی دسته‌بندی نوع میوه (سیب، موز، پرتقال)، هیچ‌گونه رابطه‌ی عددی یا ترتیبی بین کلاس‌ها وجود ندارد. اما در این مسئله، مدل باید نه‌تنها درست بودن امتیاز را بیاموزد، بلکه تفاوت بین امتیازهای نزدیک و دور را نیز در نظر بگیرد. برای مثال، پیش‌بینی ۴ به‌جای ۵ خطای کمتری نسبت به پیش‌بینی ۱ به‌جای ۵ دارد.

از آن‌جایی که چنین ساختاری با ماهیت مدل‌های رگرسیونی سازگارتر است، مدل‌هایی مانند رگرسیون خطی یا رگرسیون ترتیبی[12] برای این مسئله مناسب‌تر از مدل‌های دسته‌بندی معمول هستند. بنابراین، این مسئله به‌درستی به‌عنوان یک مسئله‌ی رگرسیون شناخته می‌شود.

● [۲.۱ (۵ نمره)] داده‌های *validation* در واقع بخشی از داده‌های آموزش ما هستند و ما از روش‌های *validation* به این خاطر استفاده می‌کنیم که پارامترهای مدل خود را به خوبی تنظیم کنیم. داده‌های *test* برای سنجیدن *generalization* مدل به کار می‌روند و فرض بر این است که ما در هنگام طراحی و انتخاب پارامترهای مدل به آن‌ها دسترسی نداریم؛ بنابراین اگر از آن‌ها برای *validation* استفاده کنیم نمی‌توانند سنجش *generalization* مدل را به درستی انجام دهند. در حالت کلی داده‌هایی که کنار گذاشته‌ایم تخمین‌گیری ناریب هستند، اما اگر با استفاده از آن‌ها *stopping early* انجام دهیم یا پارامترهای مدل را انتخاب کنیم (که در *validation cross K-fold* این کار را انجام می‌دهیم)، حتی با وجود این که مدل را روی آن‌ها آموزش نداده‌ایم دچار *bias* شده و تخمین نسبتا بهتری از واقعیت ارائه می‌دهیم.

● [۳.۱ (۵ نمره)] خطای تجربی هر دو برابر خواهد بود. فرض کنید که مدل h_1 خطای تجربی کمتری داشته باشد و w_1^*, w_2^* پارامترهای متناظر با خطای تجربی کمینه باشند. حال اگر $w = w_1^{*2} + w_2^*$ قرار دهیم، آنگاه خطای h_2 نیز همین مقدار خواهد بود و دو مدل دقیقا یکسان می‌شوند. بالعکس فرض کنید که خطای تجربی h_2 کمتر باشد و w^* پارامتر متناظر با خطای تجربی کمینه باشند. حال اگر $w_1 = 0, w_2 = w^*$ قرار دهیم، آنگاه خطای h_1 نیز همین مقدار خواهد بود و دو مدل دقیقا یکسان می‌شوند. پس دو مدل خطای تجربی یکسانی خواهند داشت.

● [۴.۱ (۵ نمره)] می‌خواهیم که اگر $\sum_{i=1}^n x_i \leq \frac{n-1}{2}$ خروجی یک بدهد و در غیر این‌صورت صفر. بنابراین اگر رابطه‌ی قبل را به فرم $\frac{n-1}{2} - \sum_{i=1}^n x_i \geq 0$ بنویسیم، آنگاه کافی است که

$$\frac{n-1}{2} + \sum_{i=1}^{n} (-1) x_i \geq 0$$

باشد که در این حالت ضرایب را می‌توانیم به صورت $w_0 = \frac{n-1}{2}$ and $w_1 = \cdots = w_n = -1$ قرار دهیم.

● [۵.۱ (۵ نمره)] در مقادیر بسیار پایین برای c خط اریبی خواهیم داشت که به درستی داده‌ها را از یکدیگر جدا خواهد کرد اما با افزایش مقدار c خط جداکننده به مرور به محور افقی تبدیل خواهد شد که در این حالت نخست w_1 خواهد شد و پس از آن w_2 به صفر میل خواهد کرد.

سوال ۲

با توجه به این که در این جا تعداد داده‌هایی که داریم محدود است و از طرفی با تعدد ویژگی‌ها مواجه هستیم، از رویکرد naive bayes استفاده می‌کنیم که این مسئله را از دیدگاه احتمالاتی بررسی می‌کند

[11] classification
[12] ordinal regression

To predict the class label for the above attribute set, we will first calculate the probability of the species being M or H in total.

$$P(\text{Species} = M) = \frac{4}{8} = 0.5$$

$$P(\text{Species} = H) = \frac{4}{8} = 0.5$$

Next, we will calculate the conditional probability of each attribute value for each class label.

$$P(\text{Color} = \text{White} \mid \text{Species} = M) = \frac{2}{4} = 0.5$$

$$P(\text{Color} = \text{White} \mid \text{Species} = H) = \frac{3}{4} = 0.75$$

$$P(\text{Color} = \text{Green} \mid \text{Species} = M) = \frac{2}{4} = 0.5$$

$$P(\text{Color} = \text{Green} \mid \text{Species} = H) = \frac{1}{4} = 0.25$$

$$P(\text{Legs} = 2 \mid \text{Species} = M) = \frac{1}{4} = 0.25$$

$$P(\text{Legs} = 2 \mid \text{Species} = H) = \frac{4}{4} = 1$$

$$P(\text{Legs} = 3 \mid \text{Species} = M) = \frac{3}{4} = 0.75$$

$$P(\text{Legs} = 3 \mid \text{Species} = H) = \frac{0}{4} = 0$$

$$P(\text{Height} = \text{Tall} \mid \text{Species} = M) = \frac{3}{4} = 0.75$$

$$P(\text{Height} = \text{Tall} \mid \text{Species} = H) = \frac{2}{4} = 0.5$$

$$P(\text{Height} = \text{Short} \mid \text{Species} = M) = \frac{1}{4} = 0.25$$

$$P(\text{Height} = \text{Short} \mid \text{Species} = H) = \frac{2}{4} = 0.5$$

$$P(\text{Smelly} = \text{Yes} \mid \text{Species} = M) = \frac{3}{4} = 0.75$$

$$P(\text{Smelly} = \text{Yes} \mid \text{Species} = H) = \frac{1}{4} = 0.25$$

$$P(\text{Smelly} = \text{No} \mid \text{Species} = M) = \frac{1}{4} = 0.25$$

$$P(\text{Smelly} = \text{No} \mid \text{Species} = H) = \frac{3}{4} = 0.75$$

Now that we have calculated the conditional probabilities, we will use them to calculate the probability of the new attribute set belonging to a single class.

Let us consider $X = \{\text{Color} = \text{Green}, \text{Legs} = 2, \text{Height} = \text{Tall}, \text{Smelly} = \text{No}\}$.

Then, the probability of X belonging to Species M will be as follows:

$$P(M \mid X) = \frac{0.5 \times 0.5 \times 0.25 \times 0.75 \times 0.25}{P(X)}$$

$$= \frac{0.0117}{P(X)}$$

Similarly, the probability of X belonging to Species H will be calculated as follows:

$$P(H \mid X) = \frac{0.5 \times 0.25 \times 1 \times 0.5 \times 0.75}{P(X)}$$

$$= \frac{0.0468}{P(X)}$$

So we will assign the entity X to species H.

سوال ۳

$$P(x|\theta) = \theta e^{-\theta x} \qquad \theta > 0,\ x \geq 0$$

(4)

.۱

$$L(x,\theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n e^{-\theta \sum_{i=1}^{n} x_i} \Rightarrow \ln L = n \ln \theta - \theta \sum_{i=1}^{n} x_i$$

$$\frac{\partial \ln L}{\partial \theta} = \frac{n}{\theta} - \sum_{i=1}^{n} x_i \Rightarrow \frac{n}{\hat{\theta}} - \sum_{i=1}^{n} x_i = 0 \Rightarrow \frac{1}{\hat{\theta}} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\Rightarrow \boxed{\hat{\theta} = \frac{1}{\frac{1}{n} \sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}}$$

شکل ۱:

$$P(\theta|D,\hat{\lambda}) = \frac{P(D|\theta)P(\theta|\hat{\lambda})}{P(D)} = \frac{P(D|\theta)P(\theta|\hat{\lambda})}{\int P(D|\theta)P(\theta|\hat{\lambda})d\theta} \propto P(D|\theta)P(\theta|\hat{\lambda})$$

$$P(\theta|\hat{\lambda}) = e^{-\hat{\lambda}\theta}, \quad P(D|\theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n e^{-\theta \sum_{i=1}^{n} x_i}$$

$$\Rightarrow P(\theta|D,\hat{\lambda}) \propto \theta^n e^{-(\hat{\lambda} + \sum_{i=1}^{n} x_i)\theta} \Rightarrow \boxed{p(\theta|D,\hat{\lambda}) = \text{Gamma}(\theta \mid n+1,\ \hat{\lambda} + \sum_{i=1}^{n} x_i)}$$

شکل ۲:

سوال ۴

$$w^* = \arg\min_w J_\lambda(w), \qquad J_\lambda(w) = \frac{1}{2}\|y - Xw\|_2^2 + \lambda\|w\|_1 \quad (\lambda > 0)$$

$$J_\lambda(w) = \frac{1}{2}(y - Xw)^T(y - Xw) + \lambda\|w\|_1$$

$$= \frac{1}{2}(y^Ty + w^TX^TXw - 2(Xw)^Ty) + \lambda\|w\|_1$$

$$(X^TX = I)$$

$$= \frac{1}{2}y^Ty + \frac{1}{2}w^Tw - (Xw)^Ty + \lambda\|w\|_1$$

$$Xw = \sum_{i=1}^n w_i X_i \Rightarrow (Xw)^Ty = \left(\sum_{i=1}^n w_i X_i\right)^T y = \sum_{i=1}^n (w_i X_i)^T y$$

$$\Rightarrow J_\lambda(w) = \frac{1}{2}y^Ty + \frac{1}{2}\sum_{i=1}^n w_i^2 - \sum_{i=1}^n (w_i X_i)^T y + \lambda\sum_{i=1}^n |w_i|$$

$$= \underbrace{\frac{1}{2}y^Ty}_{g(y)} + \sum_{i=1}^n \overbrace{\frac{1}{2}w_i^2 - w_i X_i^T y + \lambda|w_i|}^{f(X_i, y, w_i, \lambda)}$$

$$\Rightarrow J_\lambda(w) = g(y) + \sum_{i=1}^n f(X_i, y, w_i, \lambda)$$

$$\Rightarrow \frac{\partial J}{\partial w_i} = \frac{\partial f(X_i, y, w_i, \lambda)}{\partial w_i} = 0 \Rightarrow w_i^* = h(X_i, y, \lambda)$$

در نتیجه w_i^* را می‌توان تنها از X_i (ویژگی ام) و y (خروجی) تعیین کرد. بنابراین X_i ها مستقل نیز هستند.

شکل ۳:

$$J_\lambda(w) = \frac{1}{2} y^T y + \sum_{i=1}^{n} \frac{1}{2} w_i^2 - w_i X_i^T y + \lambda |w_i|$$

$$\frac{\partial J}{\partial w_i} = w_i - X_i^T y + \lambda \, sgn(w_i)$$

$$\Rightarrow w_i^* - X_i^T y + \lambda \, sgn(w_i^*) = 0$$

$$\Rightarrow \boxed{w_i^* = X_i^T y - sgn(w_i^*)\lambda}$$

۲. $w_i^* > 0$

$$\boxed{w_i^* = X_i^T y - \lambda}$$

۳. $w_i^* < 0$

$$\boxed{w_i^* = X_i^T y + \lambda}$$

۴.

$$w_i^* = 0 \Rightarrow X_i^T y = sgn(w_i^*)\lambda \Rightarrow \boxed{\lambda = -sgn(w_i^*) X_i^T y}$$

اگر مقدار λ را برابر عبارت فوق قرار دهیم، w_i^* صفر خواهد شد.

شکل ۴:

سوال ۵

الف) با بسط داده شده برای خطای تعمیم‌پذیری داریم:

$$E[(y - \hat{y})^2] = E[(f(x) + \varepsilon - \hat{y})^2] = E[(f(x) - \hat{y})^2] + 2E[(f(x) - \hat{y})\varepsilon] + E[\varepsilon^2]$$

از آن‌جایی که $E[\varepsilon] = 0$ و مستقل بودن، داریم:

$$E[(f(x) - \hat{y})\varepsilon] = 0$$

در نتیجه:

$$E[(y-\hat{y})^2] = E[(f(x)-\hat{y})^2] + \sigma_\varepsilon^2$$

اکنون عبارت اول را گسترش می‌دهیم:

$$E[(f(x)-\hat{y})^2] = E[(f(x)-E[\hat{y}]+E[\hat{y}]-\hat{y})^2]$$

$$= E[(f(x)-E[\hat{y}])^2] + E[(E[\hat{y}]-\hat{y})^2] + 2E[(f(x)-E[\hat{y}])(E[\hat{y}]-\hat{y})]$$

از آن‌جایی که جمله‌ی سوم صفر است، خواهیم داشت:

$$E[(y-\hat{y})^2] = (f(x)-E[\hat{y}])^2 + \mathrm{Var}(\hat{y}) + \sigma_\varepsilon^2$$

که در آن عبارت اول همان Bias^2 است، عبارت دوم Variance مدل، و عبارت سوم Noise یا خطای کاهش‌ناپذیر داده‌ها است.

در نتیجه خطای تعمیم‌پذیری به صورت مجموع سه مؤلفه نوشته می‌شود. این عبارت نشان‌دهنده‌ی Bias-Variance Tradeoff در مدل است. در واقع، وقتی مدل بسیار ساده باشد، Bias بالا و Variance پایین است. برعکس، اگر مدل بیش از حد پیچیده باشد، Bias کاهش یافته ولی Variance افزایش می‌یابد. هدف در یادگیری ماشین یافتن تعادلی مناسب بین این دو است.

ب) در اثر استفاده از رگولاریزاسیون L2 در مدل رگرسیون خطی با اضافه کردن جمله‌ای شامل مجموع مربعات ضرایب به تابع هزینه، از بزرگ شدن بیش از حد ضرایب جلوگیری شده و در نتیجه پیچیدگی آن کنترل می‌شود. تأثیر اصلی این روش در کاهش واریانس مدل است، زیرا مدل کمتر به نوسانات داده‌های آموزشی حساس می‌شود و در نتیجه پیش‌بینی‌های پایدارتری ارائه می‌دهد، هرچند این کار ممکن است با افزایش کمی در بایاس همراه باشد چون مدل انعطاف‌پذیری خود را محدود می‌کند. با این حال عموما افزایش بایاس نسبتاً جزئی است و کاهش قابل‌توجه واریانس منجر به بهبود خطای تعمیم‌پذیری کلی می‌شود، به‌ویژه زمانی که مدل دچار overfitting باشد.

پ) آموزش مدل بر روی یک مجموعه داده کوچکتر واریانس را افزایش می‌دهد. زیرا مدل حساسیت بیشتری به مجموعه داده های آموزشی محدود دارد. با تعداد داده کمتر ممکن است مدل نویز موجود در داده‌ها را بیشتر برازش کند. از طرفی بایاس مدل معمولا با در نظر گرفتن زیرمجموعه‌ای از داده‌ها تغییر چشمگیری نمی‌کند. زیرا بایاس بیشتر به ساختار مدل (مثلا خطی بودن یا غیرخطی بودن آن) وابسته است. بدین ترتیب از رابطه بخش آ می‌توان نتیجه گرفت با افزایش واریانس و تقریبا ثابت ماندن بایاس، خطای تعمیم‌دهی نیز افزایش می‌یابد.

ت) با افزایش درجه چند جمله‌ای مدل پیچیده‌تر و انعطاف پذیرتر می‌شود. در نتیجه مدل می‌تواند داده‌های آموزش را به طور دقیق‌تر برازش کند. بنابراین بایاس کاهش می‌یابد. از سوی دیگر، با افزایش درجه چند جمله‌ای، مدل به شدت به تغییرات جزئی داده حساس می‌شود و تغییرات کوچک داده‌ها می‌تواند به تغییرات بزرگ در پیشبینی مدل منجر شود که این مساله باعث افزایش واریانس می‌شود.