دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده مهندسی برق و کامپیوتر

یادگیری ماشین

تمرین شماره ۲

نام و نام خانوادگی
مرتضی ملکی‌نژاد شوشتری

شماره دانشجویی
۸۱۰۱۰۴۲۵۶

۲۶ آذر ۱۴۰۴

فهرست مطالب

۱ چکیده ۱

۱ Window Parzen ۲

۱.۱ . . . . . ۲

۲.۱ . . . . . ۳

۳.۱ . . . . . ۳

۴.۱ . . . . . ۴

۲ Spectrum Eigenvalue the of Shift ۵

۱.۲ . . . . . ۵

۲.۲ . . . . . ۵

۳ Kernel a Optimizing ۷

۱.۳ . . . . . ۷

۲.۳ . . . . . ۷

۴ Dimensionality of Curse the and KNN ۷

۵ Metrics Choosing ۸

۶ Regression Ridge Kernel ۹

۷ Regression Linear Deriving ۱۰

۸ Lasso of Interpretation Bayesian A ۱۱

۹ Kernel Gaussian with Classification Window Parzen ۱۲

۱۰ estimation likelihood Maximum

۱۳

فهرست تصاویر

۱ نمودار $\hat{p}$ به ازای مقادیر مختلف h ۳

۲ مقدار $\hat{p}$ به ازای $a = 1, h = \frac{1}{1.98}$ در بازه $[0, 0.04]$ ۴

۳ نمودار پراکندگی نقاط ۱۳

۴ میزان likelihood به ازای مقادیر مختلف میانگین و واریانس. مقادیر محاسبه شده به صورت تئوری به صورت خط به نمایش در آمده‌اند. ۱۴

فهرست جداول

۱ احتمال تخمین زده شده برای هر نقطه با $h = 1$ ۱۲

۲ احتمال تخمین زده شده برای هر نقطه با $h = 0.1$ ۱۲

چکیده

این تمرین برای تسلط روی Parzen ، KNN ، Hypothesis Testing و توزیع χ^2 و توزیع‌های مشتق از آن و استفاده‌های آن (مثلا توزیع واریانس) ارائه شده است.

Parzen Window

۱.۱

فرمول پارزن برابر است با:

$$\hat{P}(X) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{h^d}\phi(\frac{x-x_i}{h})$$

که در این‌جا d برابر یک است زیرا فقط یک بعد داریم. چون هدف بدست آوردن میانگین $\hat{P}(X)$ می‌باشد، باید امید را به شیوه‌ای حساب کرد که در آن داده متغیر تصادفی باشد. (که در فرمول نهایی x موجود باشد.) چون داده‌ها نیز از توزیع جامعه هستند، می‌توان احتمال را جایگذاری کرد. همچنین چون در فرمول parzen داریم میانگین n متغیر تصادفی را حساب می‌کنیم و می‌دانیم که امید یک متغیر iid با امید میانگین همه برابر است می‌توان صرفا برای یک متغیر تصادفی حساب کرد. پس داریم:

$$E(\hat{P}(Y)) = \int_0^a \frac{1}{h^d}\phi(\frac{x-y}{h})\frac{1}{a}d_y = \frac{1}{ha}\int_0^a \phi(\frac{x-y}{h})d_y$$

اگر x از y کمتر باشد مقدار ϕ صفر می‌شود. پس اگر x از صفر کمتر باشد، مقدار داخل انتگرال همواره صفر است و خروجی صفر می‌شود.

اگر x بین صفر و a باشد داریم:

$$E(\hat{P}(Y)|0 < x \leq a) = \frac{1}{ha}\int_0^x \phi(\frac{x-y}{h})d_y = \frac{1}{ha}\int_0^x e^{\frac{y-x}{h}}$$

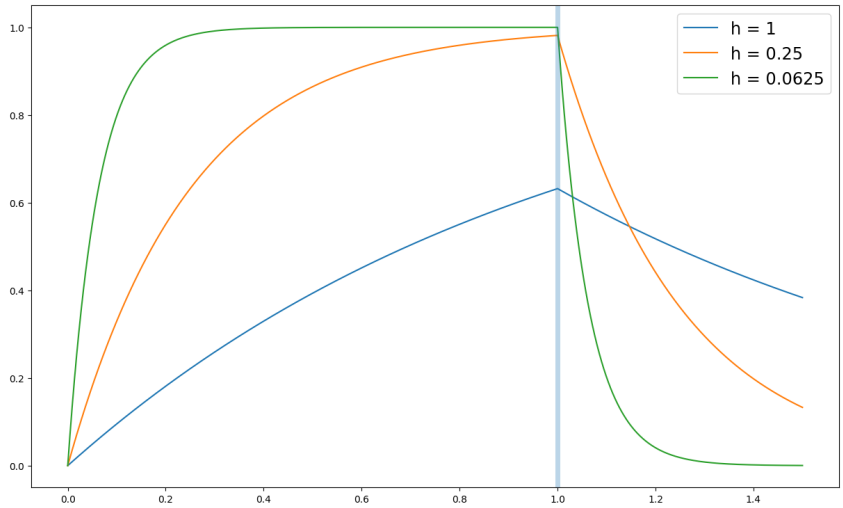
$$= \frac{1}{ha}[he^{\frac{y-x}{h}}]_0^x = \frac{1}{a}(1-e^{\frac{-x}{h}})$$

اگر x از a بزرگ‌تر باشد داریم:

$$E(\hat{P}(Y)|0 < x \leq a) = \frac{1}{ha}\int_0^a \phi(\frac{x-y}{h})d_y = \frac{1}{ha}\int_0^a e^{\frac{y-x}{h}}$$

$$= \frac{1}{ha}[he^{\frac{y-x}{h}}]_0^a = \frac{1}{a}(e^{\frac{a-x}{h}} - e^{\frac{-x}{h}}) = \frac{1}{a}(e^{\frac{a}{h}-1})e^{\frac{-x}{h}}$$

۲.۱



شکل ۱ : نمودار $\hat{p}$ به ازای مقادیر مختلف h

۳.۱

داریم:

$$\int_0^a (\hat{p}(x) - p(x))d_x = \int_0^a \frac{1}{a}(e^{\frac{-x}{h}}) = \frac{1}{a}[-h\, e^{\frac{-x}{h}}]_0^a = \frac{-h}{a}[e^{\frac{-x}{h}}]_0^a =$$

$$\frac{h}{a}(1 - e^{\frac{-a}{h}})$$

با جایگذاری متغیر $t = \frac{-a}{h}$ داریم:

$$\frac{1-e^t}{-t} \leq 0.01 \Rightarrow \frac{e^t - 1}{t} \leq 0.01$$

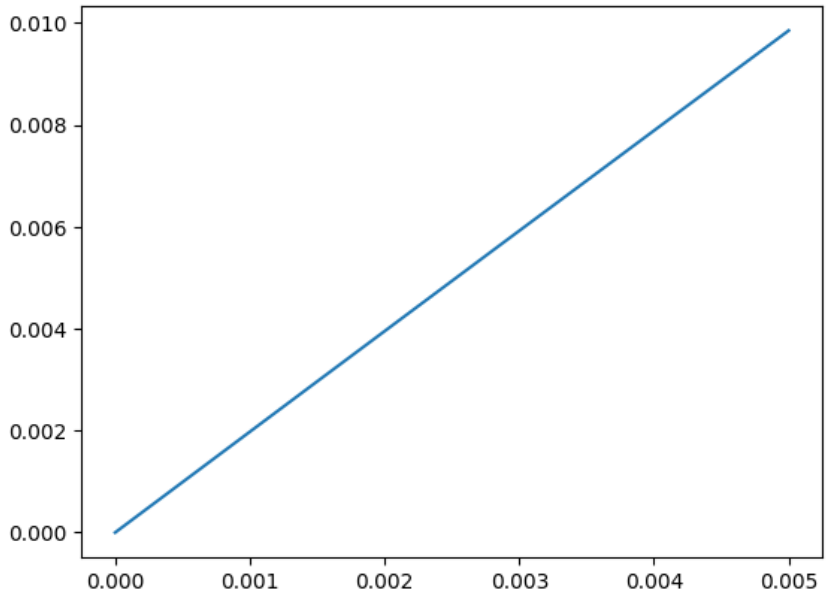
اگر بسط تیلور $e^t = 1 + t + \frac{t^2}{2}$ را بنویسیم داریم:

$$1+0.5t \leq 0.01 \Rightarrow 0.5t \leq 0.99 \Rightarrow t \leq 1.98 \Rightarrow \frac{a}{h} \geq 1.98 \Rightarrow h \leq \frac{a}{1.98}$$

یعنی برای بایاس کمتر از یک درصد، h باید کمتر از حدودا نصف a باشد.

۴.۱

$$a = 1 \Rightarrow h = \frac{1}{1.98} \approx 0.5$$



شکل ۲: مقدار $\hat{p}$ به ازای $a = 1, h = \frac{1}{1.98}$ در بازه $[0, 0.04]$

Shift of the Eigenvalue Spectrum

$$\xi A = \lambda_A I, \xi B = \lambda_B I \Rightarrow \xi(A+B) = (\lambda_A + \lambda_B)I$$

یعنی ξ بردار ویژه $A+B$ است و مقدار ویژه آن برابر با جمع دو مقدار ویژه می‌باشد.

۱.۲

حالت تکین وقتی پیش می‌آید که دترمینان صفر باشد یا به عبارت دیگر، یکی از مقادیر ویژه ماتریس X^TX صفر باشد. واضح است که با جمع کردن ماتریس با λI مقدار ویژه‌های ماتریس با λ جمع می‌شود

$$U = x^TX \quad \xi U = \gamma I \Rightarrow \xi(U+\lambda I) = \xi U + \xi\lambda = \gamma\xi + \lambda\xi = (\gamma+\lambda)\xi \Rightarrow$$
$$\xi(U+\lambda I) = (\gamma+\lambda)\xi$$

که یعنی با جمع λI مقادیر ویژه بردار U با λ جمع می‌شوند. از آنجایی که $U = X^TX$ یک ماتریس semi positive definite است، مقادیر ویژه آن همه نامنفی هستند. پس اگر λ یک مقدار مثبت باشد، نتیجه یک ماتریس positive definite است که دترمینان آن مثبت است و جواب دارد.

۲.۲

linear $\Rightarrow \hat{\beta} = argmin(y-\beta X)^2 = (X^TX)^{-1}X^Ty$

Ridge $\Rightarrow \hat{\beta} = argmin(y-\beta X)^2 + (\lambda\beta)^2 = (X^TX+\lambda I)^{-1}X^Ty$

واضح است که در ridge مقدار bias بیشتر است زیرا بعد از فیت شدن نیز به علت وجود λ مدل روی داده کامل فیت نمی‌شود. در مقابل در linear regression چون هدف فیت شدن کامل روی داده است بایاسی اضافه نمی‌شود و بایاس موجود ناشی از داده است.

در ridge چون یکی از اهداف کم بودن مقدار β می‌باشد، احتمال بیش‌برازش کمتر است.

برای مثال ممکن است مدل به دو ویژگی در مقابل هم، امتیاز های قرینه ولی بزرگی بدهد که روی داده آموزش به طور کامل خنثی شوند ولی موقع تست چون ممکن است مقدار این ویژگی ها اندکی باهم تفاوت داشته باشد، منجر به خطای بالایی خواهد شد. ولی در ridge با جلوگیری از زیاد شدن β جلوی این مورد گرفته می‌شود.

برای جمع بندی، ridge کمی bias بیشتری دارد ولی variance کمتری دارد. این tradeoff را می‌توان با مقدار λ کنترل کرد.

Optimizing a Kernel

۱.۳

$$L(x_i) = \sum_i (y_i^2 + \hat{y_i}^2 - 2y_i\hat{y}_i) = (y_i - (\frac{\sum_j K(x_i, x_j)y_i}{\sum_j K(x_i, x_j)}))^2$$

اگر برای $L(x_i)$ مقدار خود $x_j = x_i$ نیز درنظر گرفته شود، احتمال بیش‌برازش وجود دارد پس بهتر است که خود متغیر درنظر گرفته نشود پس داریم:

$$L = \sum_i (y_i^2 + \hat{y_i}^2 - 2y_i\hat{y}_i) = \sum_i (y_i - (\frac{\sum_{j\neq i} K(x_i, x_j)y_i}{\sum_{j\neq i} K(x_i, x_j)}))^2$$

$$= \sum_i (y_i - (\frac{\sum_{j\neq i} e^{-(x_i - x_j)W(x_j - xI)^T} y_i}{\sum_{j\neq i} e^{-(x_i - x_j)W(x_j - xI)^T}}))^2$$

۲.۳

KNN and the Curse of Dimensionality

a

اگر بازه $[a, b]$ انتخاب شود باید مقدار $P(a \leq x \leq b)$ حساب شود. در توزیع یکنواخت این عبارت برابر است با طول بازه تقسیم بر طول کل که در این حالت خاص برابر با 0.1 می‌باشد.

b

باتوجه به اینکه دو ویژگی از هم مستقل هستند احتمال آن‌ها در هم ضرب می‌شود یعنی برابر است با 0.01

c

مشابه قسمت قبل با ضرب احتمال‌ها داریم:

$$P = 0.1^p = 0.1^1 00$$

d

اگر تعداد داده نسبت به ابعاد بصورت نمایی رشد نکند، برای تخمین هر نقطه داده کمی موجود است و نمی‌توان با دقت خوبی پیش‌بینی کرد.

در روش parzen این مشکل به این صورت رخ پیدا می‌کند که جاهای زیادی در pdf estimate صفر می‌باشد (زیرا داده‌ای یافت نشده است.) در روش KNN این مشکل به اینصورت رخ می‌دهد که پنجره بزرگ و بزرگ تر می‌شود و احتمالا از داده‌های دور استفاده شود که نتیجتا تخمین خوبی نمی‌دهد.

Metrics Choosing

a

استفاده از cosine بهتر است زیرا در مقادیری که ذکر شد،نسبت کافئین به شکر (یا برعکس) جداکننده بهتری نسبت به شکر یا کافئین به تنهایی است.

b

در اینجا چون بردار صفر وارد می‌شود نسبت کافی نیست (اگر صفر هم نبود احتمال اشتباه گرفتن آب با انرژی زا وجود داشت زیرا نسبت تقریبا یکسانی دارند.) در اینجا استفاده از فاصله اقلیدسی منطقی است زیرا مقدار عددی ویژگی‌ها اهمیت دارد.

c

چون مقدار ها عددی نیستند فاصله همینگ بهتر است.

d

گزینه ۳. در درخت تصمیم، میزان تاثیر ویژگی‌ها سنجیده می‌شود ولی در سنجش فاصله این عدد سنجیده نمی‌شود.

Kernel Ridge Regression

a

با استفاده از vector notation داریم:

$$J = \frac{1}{2}\sum(X\theta - Y)^2 + \frac{\lambda}{2}|\theta|^2 = \frac{1}{2}(X\theta - Y)^T(X\theta - Y) + \frac{\lambda}{2}\theta^T\theta$$

$$= \frac{1}{2}(\theta^T X^T X\theta - \theta^T X^T Y - Y^T X\theta + Y^T Y + \lambda\theta^T\theta)$$

این عبارات عدد هستند پس ترانهاد آن‌ها باهم برابر است. دوعبارت وسط ترانهاد همدیگر هستند و باهم برابر هستند پس داریم:

$$J = \frac{1}{2}(\theta^T X^T X\theta^T - 2\theta^T X^T Y + Y^T Y + \lambda\theta^T\theta)$$

با مشتق گرفتن نسبت به θ و صفر گذاشتن داریم:

$$\frac{1}{2}(2X^T X\theta - 2X^T Y + 2\lambda\theta) = 0 \Rightarrow (X^T X + \lambda I)\theta = X^T Y$$
$$\Rightarrow \theta = ((X^T X + \lambda I)^{-1} X^T Y)$$

b

حل نشده است.

Deriving Linear Regression

a

داریم:

$$E(yx) = w^T x^2 + \epsilon x = w^T E(x^2) + E(\epsilon)E(x) = w^T \Sigma_x$$

b

می‌دانیم بهترین تخمینگر میانگین جامعه، میانگین نمونه است. پس داریم:

$$E(yx|D) = \frac{1}{n}\sum_{i=1}^{n}(x_i y_i)$$

c

داریم:

$$L = E((w^T x - y)^2) \xrightarrow{minimize\ loss \rightarrow derviative\ from\ w} E(2(w^T x - y)x) = 0 \Rightarrow$$

$$E(w^T x^2) = E(yx) \Rightarrow w^T \Sigma_x \Rightarrow \hat{w} = (\frac{E(yx)}{\Sigma_X})^T = (\frac{\Sigma_{xy}}{\Sigma_x})^T$$

d

روی صورت اتفاقی نمی‌افتد زیرا مقدار آن از روی نمونه حساب شد. ولی در مخرج، درواقع $E(X^2)$ بود. پس یعنی باید با مربع میانگین جمع شود پس داریم:

$$\hat{w} = \frac{\Sigma_{xy}}{\Sigma_x + \mu_x^2}$$

e

بله. درجایی از محاسبات از نرمال بودن استفاده نشد. صرفا اگر نرمال باشد می‌توان چیزهایی مثل بازه اطمینان و غیره را حساب کرد.

A Bayesian Interpretation of Lasso

a

داریم:

$$f(w|(y|x)) = \frac{f(y|x,w)f(w)}{f(y|x)}$$

چون مقدار w در مخرج نقشی ندارد و برای همه w ها یکسان است می‌توان از آن صرفنظر کرد. پس یعنی باید w ای پیدا کرد که مقدار $f(w)f(y|x,w)$ در آن بیشینه باشد. پس داریم:

$$f(w|(y|x)) \propto f(w)f(y|x,w) = \frac{1}{2b}e^{\frac{-|w|}{b}} * \prod \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(y-w.x)^2}{2\sigma^2}}$$

با حذف مقادیری که w در آن‌ها نقشی ندارد داریم:

$$f(w|(y|x)) \propto e^{-(\frac{|w|}{b}+\sum\frac{(y-w.x)^2}{2\sigma^2})}$$

b

با گرفتن لگاریتم از حاصل قسمت قبل داریم :

$$l = -(\frac{|w|}{b}+\sum\frac{(y-w.x)^2}{2\sigma^2})$$

از آنجایی که لگاریتم صعودی است کافیست مقدار l بیشینه شود تا maximum likelihood بدست آید. همچنین می‌توان یک عدد ثابت را ضرب کرد پس داریم:

$$\hat{w} = argmax\frac{1}{2b\sigma^2}|w| + \sum(y-w.x)^2$$

که این همان چیزی است که در سوال آمده و مقدار λ برابر است با:

$$\frac{1}{2b\sigma^2}$$

Parzen Window Classification with Gaussian Kernel

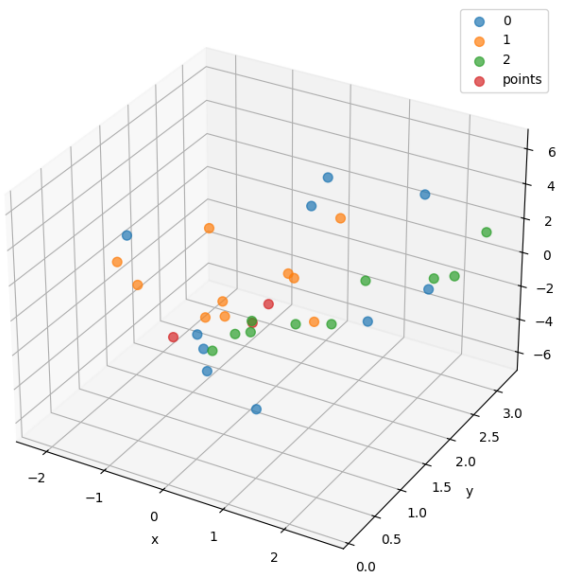
موارد در نوتبوک 10.ipynb حل شده‌اند. در هر دو مورد کلاس ۱ انتخاب شد.

جدول ۱: احتمال تخمین زده شده برای هر نقطه با $h = 1$

Coordinates	$\mathbf{P}(\omega_1)$	$\mathbf{P}(\omega_2)$	$\mathbf{P}(\omega_3)$
$(0.5, 1.0, 0.0)$ 1	0.1259	0.4711	0.3775
$(0.31, 1.51, -0.5)$	0.1534	0.4828	0.2050
$(-0.3, 0.44, -0.1)$	0.1399	0.3783	0.1771

جدول ۲: احتمال تخمین زده شده برای هر نقطه با $h = 0.1$

Coordinates	$\mathbf{P}(\omega_1)$	$\mathbf{P}(\omega_2)$	$\mathbf{P}(\omega_3)$
$(0.5, 1.0, 0.0)$	8.77×10^{-20}	6.77×10^{-5}	8.00×10^{-17}
$(0.31, 1.51, -0.5)$	2.87×10^{-20}	1.20×10^{-5}	2.16×10^{-25}
$(-0.3, 0.44, -0.1)$	3.64×10^{-12}	3.78×10^{-4}	1.06×10^{-39}



شکل ۳: نمودار پراکندگی نقاط

Maximum likelihood estimation

a,b

$$f(x|\mu,\sigma) \propto \frac{1}{\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

برای تخمین mle باید این عبارت را براساس μ و σ بیشینه کرد. برای میانگین واضح است که کمترین مقدار داخل پرانتز (و درنتیجه بیشترین مقدار توان) وقتی است که مقدار $\sum(x-\mu)^2$ کمینه باشد.با باز کردن این عبارت داریم:

$$\mu = argmin \sum x^2 + \mu^2 - 2x\mu = argmin(n\mu^2 - 2n\bar{x}\mu) = \mu^2 - 2\mu\bar{x} \Rightarrow \mu = \bar{x}$$

برای واریانس داریم:

$$l = \sum(-ln(\sigma) - \frac{(x-\mu)^2}{2\sigma^2})$$

با مشتق گیری برحسب σ داریم:

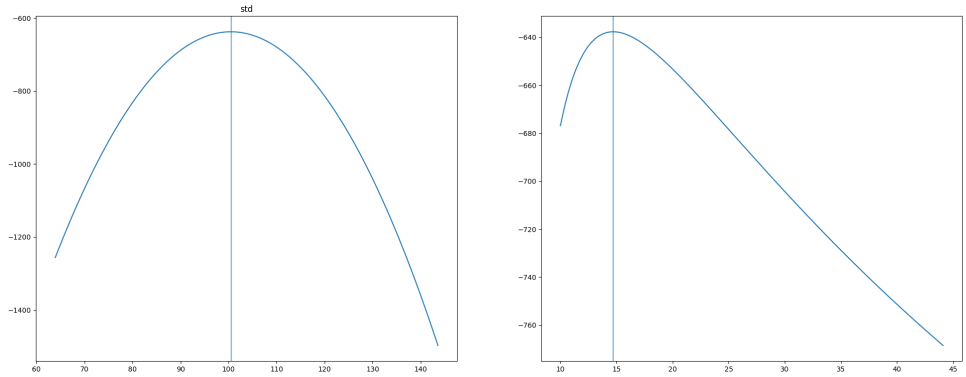
$$-(\frac{n}{\sigma}+\sum \frac{(x-\mu)^2}{2}(\frac{-2}{\sigma^3})) = \frac{-1}{\sigma}(n-\sum \frac{(x-\mu)^2}{\sigma^2}) = 0 \Rightarrow \sigma^2 = \sum \frac{(x-\mu)^2}{n}$$

که با محاسبه با اعداد بدست می‌آید:

$$\mu_{theory} = 100.48049999999999 \qquad \sigma_{theory} = 14.707769026946268$$

مقدار عددی نیز چیزی در همین حدود شد:

$$\mu_{numeric} = 100.48050398080368 \qquad \sigma_{numeric} = 14.707763577982128$$



شکل ۴: میزان likelihood به ازای مقادیر مختلف میانگین و واریانس. مقادیر محاسبه شده به صورت تئوری به صورت خط به نمایش در آمده‌اند.

c,d

سوال درواقع fisher information را می‌خواهد:

$$I^{-1} = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix} \approx \begin{bmatrix} 1.081 & 0 \\ 0 & 467.93 \end{bmatrix}$$

هم در حالت عددی و هم در حالت تئوری این‌مقدار تقریبا برابر است.

e

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

با استفاده از این فرمول بازه با استفاده از کد در نوتبوک محاسبه شد. بازه اطمینان 95 درصد انحراف معیار برابر است با:

$[13.4274, 16.3507]$

f

- برای Z test چون سایز نمونه از 30 بزرگ تر است می‌توان از خود Zاستفاده کرد. چون راجع به Alternative Hypotheis حرفی زده نشده، تست two sided درنظر گرفته می‌شود. طبق فرمول داریم:

$$z = \frac{100.48 - 103}{\frac{14.7}{\sqrt{200}}} \approx -2.42 \Rightarrow p_value = 2 * (1 - Z_{-2.42}) \approx 0.01$$

پس یعنی فرض صفر رد می‌شود.