دانشگاه آزاد اسلامی واحد تبریز

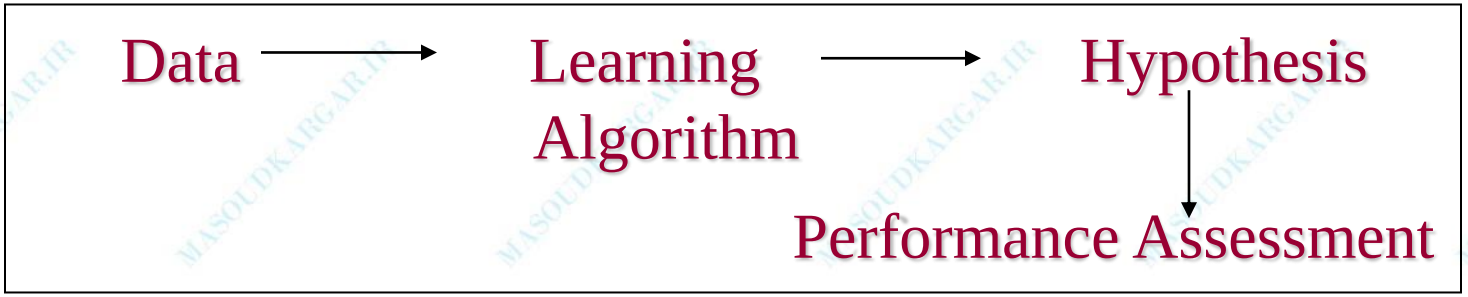
نام درس: یادگیری ماشین

بخش: ارزیابی فرضیه‌ها

نام استاد: دکتر مسعود کارگر

مقدمه

- یک الگوریتم یادگیری با استفاده از داده‌های آموزشی فرضیه‌ای را بوجود می‌آورد. قبل از استفاده از این فرضیه ممکن است که لازم شود تا دقت این فرضیه مورد ارزیابی قرار گیرد.



- اینکار از دو جهت اهمیت دارد:

1. دقت فرضیه را برای مثالهای نادیده حدس بزنیم.
2. گاهی اوقات ارزیابی فرضیه جزئی از الگوریتم یادگیری است: مثل هرس کردن درخت تصمیم.

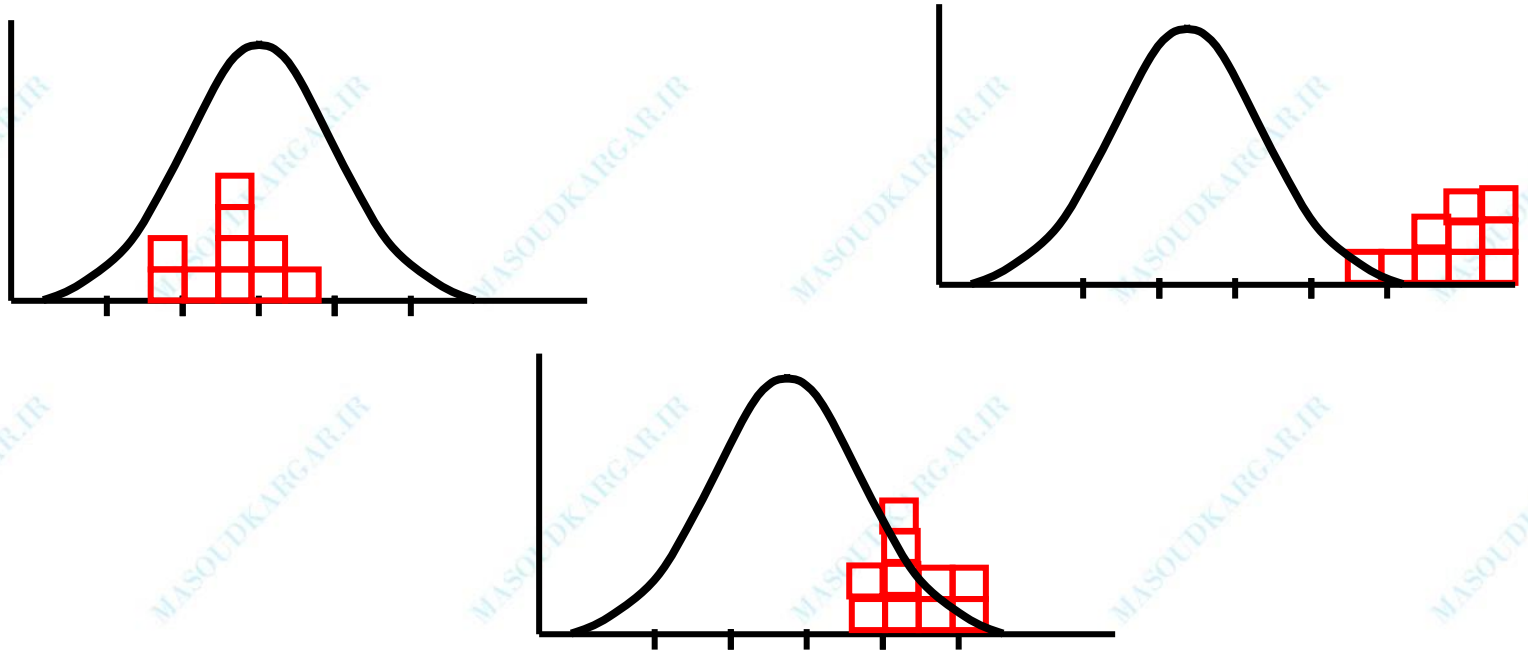
روشهای آماری

- در این فصل سعی می‌شود تا روشهای آماری مناسب برای حدس زدن دقت فرضیه‌ها معرفی گردند. مبنای کار در جهت پاسخگوئی به سه سوال زیر است:

1. اگر دقت یک فرضیه برای داده‌های محدودی معلوم باشد دقت آن برای سایر مثالها چه قدر خواهد بود؟
2. اگر یک فرضیه برای داده‌های محدودی بهتر از فرضیه دیگری عمل کند احتمال اینکه این وضعیت در حالت کلی نیز صادق باشد چقدر است؟
3. وقتی که داده آموزشی اندکی موجود باشد بهترین راه برای اینکه هم فرضیه را یاد بگیریم و هم دقت آنرا اندازه‌گیری کنیم چیست؟

آموزشی داده‌های کمی

- وقتی که داده آموزشی محدود باشد این امکان وجود دارد که این مثالها نشاندهنده توزیع کلی داده‌ها نباشند.



داده کمی مشکل

- وقتی که یادگیری با استفاده از داده‌های محدودی انجام می‌شود دو مشکل ممکن است رخ دهد:

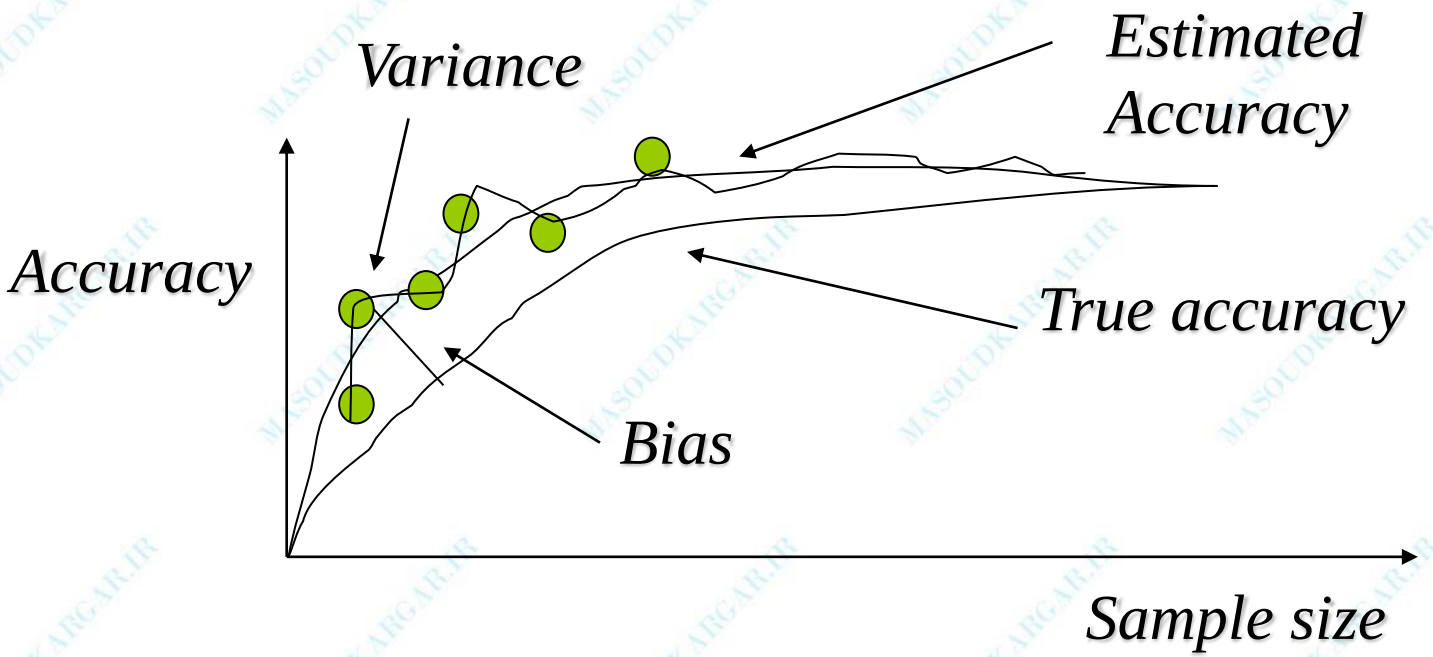
1. بایاس در تخمین.

دقت یک فرضیه بر روی مثالهای آموزشی تخمین مناسبی برای دقت آن برای مثالهای نادیده نیست. زیرا فرضیه یاد گرفته شده بر اساس این داده‌ها برای مثالهای آتی به صورت خوش‌بینانه (*optimistic*) عمل خواهد نمود. برای رهائی از این امر می‌توان از مجموعه داده‌های تست استفاده کرد.

2. انحراف (*Variance*) در تخمین.

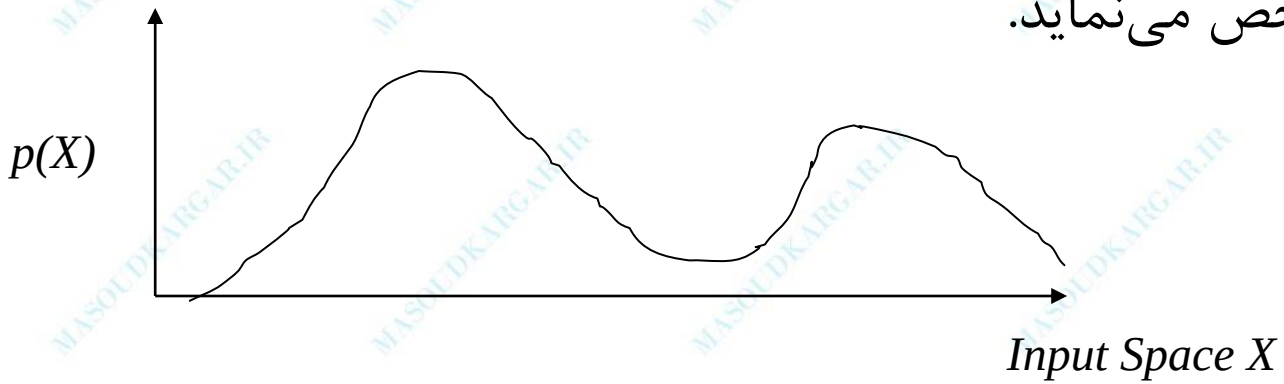
حتی با وجود استفاده از مجموعه تست این امکان وجود دارد که دقت فرضیه اندازه‌گیری شده با دقت فرضیه واقعی اختلاف داشته باشد. این دقت بستگی به خصوصیات مجموعه تست و تطابق با توزیع کلی داده‌ها دارد.

بایاس و انحراف معیار در تخمین



تخمین دقت فرضیه

- در یک مثال یادگیری می‌توان برای فضای مثالهای ورودی یک تابع توزیع احتمال نامعلوم D در نظر گرفت که احتمال رخداد هر نمونه X را با $p(X)$ مشخص می‌نماید.



- در این صورت با دو سوال زیر مواجه هستیم:

1. اگر فرضیه h و تعداد n نمونه داشته باشیم که به صورت تصادفی از مثالهائی با توزیع D انتخاب شده باشند، بهترین تخمین برای دقت h برای مثالهای آتی با همان توزیع چیست؟
2. خطای احتمالی در این تخمین دقت چقدر است؟

خطای نمونه و خطای واقعی

- خطای نمونه

خطای فرضیه روی مجموعه مثالهای موجود(آموزشی و یا تست) به عبارت دیگر کسری است از نمونه‌های S که تحت فرضیه h نسبت به تابع هدف f اشتباه دسته‌بندی شده‌اند:

$$error_S(h) = 1/n \sum_{x \in S} \delta(f(x),h(x))$$

که در ان n تعداد مثالهای S و اگر $f(x) \neq h(x)$ آنگاه مقدار $\delta(f(x),h(x))$ برابر با 1 است در غیر این صورت برابر با 0 است.

واقعی خطای و نمونه خطای

- خطای واقعی

عبارت است از خطای فرضیه روی مجموعه تمام مثالها با توزیع نامعلوم D و برابر است با احتمال اینکه یک نمونه تصادفی به اشتباه دسته‌بندی شود.

خطای واقعی فرضیه h نسبت به تابع هدف f و داده با توزیع Dبه صورت زیر بیان می‌شود:

$$error_D(h) = Pr_{x \in D}[f(x) \neq h(x)]$$

آنچه که در دست داریم خطای نمونه است در حالیکه آنچه که به دنبال آن هستیم خطای واقعی است. در این صورت باید به این سوال پاسخ دهیم که خطای نمونه تا چه حدی می‌تواند تخمین خوبی برای خطای واقعی باشد؟

مثال

- یک مجموعه داده شش‌تائی با توزیع احتمال زیر وجود دارد:

$$P(X1) = 0.2 \quad P(X4) = 0.1$$
$$P(X2) = 0.1 \quad P(X5) = 0.2$$
$$P(X3) = 0.3 \quad P(X6) = 0.1$$

فرضیه h برای مجموعه نمونه$\{X1, X2, X3, X4\}$ می‌تواند $X1, X2, X3$ را بدرستی دسته‌بندی کند ولی قادر به دسته‌بندی صحیح $X4$ نیست .دراین صورت خطای نمونه برابر است با:

$$¼ (0 + 0 + 0 + 1) = ¼ = 0.25$$

اگر این فرضیه برای $X6$ صحیح و برای$X5$ نادرست باشد در این صورت خطای واقعی برابر است با:

$$0.2(0) + 0.1(0) + 0.3(0) + 0.1(1) + 0.2(1) + 0.1(0) = 0.3$$

فاصله اطمینان برای فرضیه‌های با مقادیر گسسته

اگر سه شرط زیر برقرار باشند:

- نمونه S دارای n مثال باشد که مستقل از یکدیگر و مستقل از h برپایه توزیع احتمال D انتخاب شده باشند.
- $n \geq 30$ باشد.
- فرضیه h منجر به r خطا روی این مثالها گردد. (یعنی $error_S(h)=r/n$)

آنگاه می‌توان بر پایه قضایای آماری ادعا نمود که:

1. اگر اطلاعات بیشتری موجود نباشد، محتمل‌ترین مقدار برای $error_D(h)$ برابر با $error_S(h)$ خواهد بود
2. با احتمال 95% خطای واقعی بین فاصله زیر قرار دارد:

$$error_S(h) \pm 1.96\sqrt{\frac{error_S(h)(1-error_S(h))}{n}}$$

مثال

- فرض کنید که S دارای $n=40$ مثال بوده و فرضیه h منجر به $r=12$ خطا بر روی این داده شود. در این صورت:
- خطای نمونه برابر است با $error_S(h)=12/40=.30$
- *اگر این آزمایش را بارها و بارها برای 40 نمونه جدید تکرار کنیم متوجه خواهیم شد که در 95% مواقع خطای محاسبه شده در فاصله زیر قرار خواهد داشت:*

$$0.30 \pm (1.96 \times .07) = 0.30 \pm .14$$

گسسته مقادیر با فرضیه‌های برای اطمینان فاصله

- عبارت فوق را می‌توان بجای فاصله اطمینان 95% برای هر فاصله دیگری نظیر $N\%$ نیز ذکر نمود:

$$error_s(h) \pm Z_N \sqrt{\frac{error_s(h)(1-error_s(h))}{n}}$$

- مقدار ثابت Z_N برای درصدهای مختلف را می‌توان از جدول زیر بدست آورد:

$N\%$:	50%	68%	80%	90%	95%	98%	99%
z_N:	0.67	1.00	1.28	1.64	1.96	2.33	2.58

- این تقریب زمانی بهترین نتیجه را دارد که:

$$n\, error_S(h)(1 - error_S(h)) \geq 5$$

مقدمه‌ای بر تئوری نمونه‌برداری

- مروری بر بحثهای زیر
  - میانگین
  - واریانس
  - توزیع دوجمله‌ای
  - توزیع نرمال
  - فواصل یک طرفه و دو طرفه

خطا تخمین

- سوال: تاثیر اندازه داده‌های نمونه بر اختلاف بین خطای نمونه و خطای واقعی چیست؟
- در واقع پاسخ این سوال را متخصصین آمار داده‌اند!
- می‌توان اندازه‌گیری خطای نمونه را به آزمایشی با نتیجه تصادفی تشبیه کرد. اگر به دفعات n نمونه با توزیع احتمال D به صورت تصادفی انتخاب و خطای نمونه برای هر کدام اندازه‌گیری شود، بعلت متفاوت بودن نمونه‌ها مقدار خطا نیز متفاوت خواهد بود. نتیجه حاصل از هر آزمایش یک متغیر تصادفی خواهد بود. (متغیر تصادفی: متغیری که مقادیر مختلفی را با احتمالهایی که توسط توزیع احتمال آن مشخص می‌شود، اختیار می‌کند.)
- چنین آزمایشی را می‌توان با استفاده از توزیع دوجمله‌ای توصیف نمود.

دوجمله‌ای توزیع

- توزیع دوجمله‌ای برای آزمایشاتی استفاده می‌شود که دارای خواص زیر باشند:

1. آزمایش به تعداد n دفعه تکرار شود، n مقداری ثابت و از قبل دانسته است.
2. هر آزمایش دارای دو نتیجه درست و یا غلط باشد.
3. آزمایشات مستقل از همدیگر باشند، به نحویکه نتیجه یک آزمایش تاثیری بر سایر آزمایشات نداشته باشد.
4. احتمال وقوع نتیجه درست برای تمام آزمایشات ثابت باشد.

مثال

در پرتاب یک سکه به تعداد 8 دفعه:

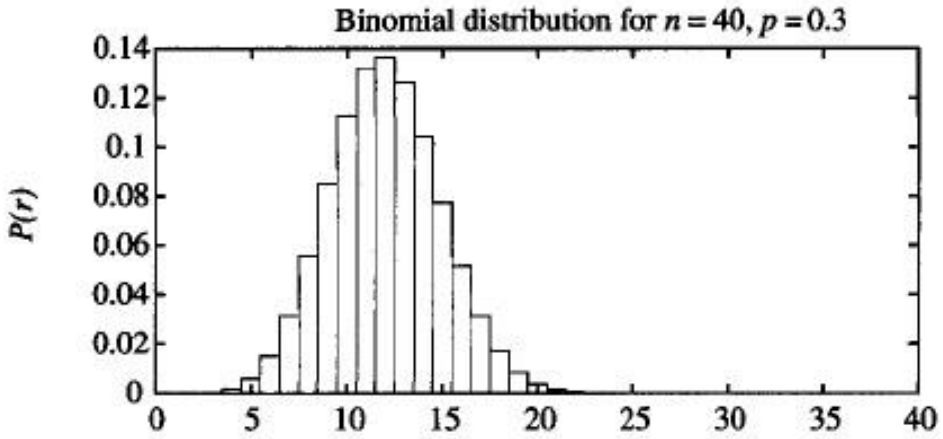
- $n=8$
- آزمایش دارای دو نتیجه شیر یا خط است.
- نتیجه هر پرتاب سکه مستقل از پرتاب‌های قبلی است.
- احتمال آمدن شیر برای هر پرتاب $p=1/2$ است.

دوجمله‌ای احتمال

- احتمال وقوع r موفقیت در N بار تکرار یک آزمایش از رابطه زیر محاسبه می‌شود:

$$P(r)=\frac{N!}{r!(N-r)!}p^{r}\ (1-p)^{N-r}$$

- که در آن p احتمال وقوع موفقیت در هر بار تکرار آزمایش است.



توزیع دوجمله‌ای برای
n=40, p=.3

مثال

- احتمال آمدن 6 خط در 8 بار پرتاب یک سکه چقدر است؟

$$p = .5$$

$$P(6) = \frac{8!}{6!(8-6)!} .5^6 \left(1-.5\right)^{8-6} = .055$$

خطای نمونه‌برداری

این خطا را می‌توان با پرتاب سکه مقایسه نمود:

- پرتاب سکه و دیدن یک خط ←انتخاب یک نمونه از D و تعیین اینکه آیا h آنرا غلط ارزیابی می‌کند یا نه.
- احتمال اینکه در یک پرتاب واحد یک خط داشته باشیم ← احتمال اینکه یک نمونه غلط ارزیابی شود.
- دیدن تعداد r خط در N بار پرتاب سکه ←تعداد ارزیابی‌های غلط از بین N نمونه انتخاب شده.

میانگین

- مقدار میانگین (*Expected Value*) برای یک متغیر تصادفی Y که می‌تواند مقادیر $y_1,\dots,y_n$ را داشته باشد عبارت است از:

$$E[Y] = \Sigma_{i=1}^{n} y_i \, Pr(Y=y_i)$$

- برای یک متغیر تصادفی با توزیع دوجمله‌ای این مقدار برابر است با:

$$E[Y] = np$$

واریانس

- واریانس گستردگی توزیع احتمال و فاصله متغیر تصادفی از مقدار میانگین را مشخص می‌کند. واریانس یک متغیر تصادفی Y عبارت است از:

$$Var[Y] = E[(Y-E[Y])^2]$$

- ریشه دوم واریانس انحراف معیار نامیده می‌شود.
- برای یک متغیر تصادفی با توزیع دوجمله‌ای این مقادیر برابراند با:

$$Var[Y] = np(1-p) \qquad \sigma_Y = \sqrt{np(1-p)}$$

تخمین‌زننده، بایاس و واریانس

- میزان اختلاف احتمالی بین خطای نمونه و واقعی چیست؟
- اگر r تعداد نمونه‌های با دسته‌بندی غلط روی مجموعه نمونه‌ی S با اندازه n برای فرضیه h باشد، در این صورت میزان خطای نمونه و واقعی با توجه به توزیع دوجمله‌ای برابر است با:

$$error_S(h) = r/n \qquad \& \qquad error_D(h) = p$$

- که p احتمال دسته‌بندی غلط یک نمونه انتخاب شده از D است.
- متخصصین آمار $error_S(h)$ را یک تخمین‌زننده $(estimator)$ برای خطای واقعی $error_D(h)$ می‌نامند.
- هر تخمین‌زننده دارای دو پارامتر مهم است: بایاس تخمین و واریانس.

بایاس تخمین

- اختلاف بین میانگین مقادیر تخمین زده شده و مقدار واقعی بایاس تخمین نامیده می‌شود.

$E[Y] - p$ (p اینجا یک پارامتر اختیاری است)

  - اگر مقدار بایاس صفر باشد، تخمین‌زننده بدون بایاس نامیده می‌شود.

- سوال: آیا خطای نمونه $error_S(h)$ یک تخمین‌زننده بدون بایاس برای خطای واقعی $error_D(h)$ می‌باشد؟
  - پاسخ این سوال مثبت است زیرا در توزیع دوجمله‌ای داریم:

$$E[r]=np \Rightarrow error_S(h) = np/n=p \quad \& \quad error_D(h) =p$$

تخمین‌زننده معیار انحراف

- هدف یافتن یک تخمین‌زننده با بایاس صفر و واریانس حداقل است.
- اگر در یک نمونه n عضوی تعداد r خطا داشته باشیم، انحراف معیار تخمین‌زننده (خطای نمونه) برابر است با

$$\sigma_{error_S(h)} = \frac{\sigma_r}{n} = \sqrt{\frac{p(1-p)}{n}}$$

- این مقدار را می‌توان به صورت زیر تقریب زد:

$$\sigma_{error_S(h)} \approx \sqrt{\frac{error_S(h)(1-error_S(h))}{n}}$$

مثال از واریانس

To illustrate these concepts, suppose we test a hypothesis and find that it commits $r = 12$ errors on a sample of $n = 40$ randomly drawn test examples. Then an unbiased estimate for $error_{\mathcal{D}}(h)$ is given by $error_S(h) = r/n = 0.3$. The variance in this estimate arises completely from the variance in r, because n is a constant. Because r is Binomially distributed, its variance is given by Equation (5.7) as $np(1-p)$. Unfortunately p is unknown, but we can substitute our estimate r/n for p. This yields an estimated variance in r of $40 \cdot 0.3(1 - 0.3) = 8.4$, or a corresponding standard deviation of $\sqrt{8.4} \approx 2.9$. This implies that the standard deviation in $error_S(h) = r/n$ is approximately $2.9/40 = .07$. To summarize, $error_S(h)$ in this case is observed to be 0.30, with a standard deviation of approximately 0.07. (See Exercise 5.1.)

فاصله اطمینان

- یک راه معمول برای توصیف عدم قطعیت یک تخمین ارایه یک محدوده مقادیر مورد انتظار برای مقدار واقعی است. این محدوده را فاصله اطمینان تخمین می‌نامند.
- با توجه به تبعیت تخمین‌زننده ما از توزیع دوجمله‌ای مقدار میانگین برابر با $error_D(h)$و مقدارانحراف معیار برابر است با

$$\sigma_{error_S(h)} \approx \sqrt{\frac{error_S(h)(1-error_S(h))}{n}}$$

- از اینرو برای بدست آوردن فاصله اطمینان 95% می‌بایست فاصله‌ای حول میانگین پیدا کینم که 95% احتمال را در بر داشته باشد.
- از آنجائیکه برای توزیع دوجمله‌ای محاسبه این مقدار مشکل بوده و از طرفی از آنجائیکه برای نمونه‌های زیاد توزیع دوجمله‌ای به توزیع نرمال نزدیک می‌شود (قضیه حد مرکزی)، می‌توان برای محاسبه فاصله اطمینان از توزیع نرمال بهره گرفت.

تقریب با توزیع نرمال

- برای توزیع نرمال با میانگین μ و واریانس σ فاصله اطمینان $N\%$ برابر است با: $\mu \pm Z_N \sigma$
- مقدار Z_N با توجه به منحنی نرمال مطابق جدول زیر است:

$N\%$:	50%	68%	80%	90%	95%	98%	99%
z_N:	0.67	1.00	1.28	1.64	1.96	2.33	2.58

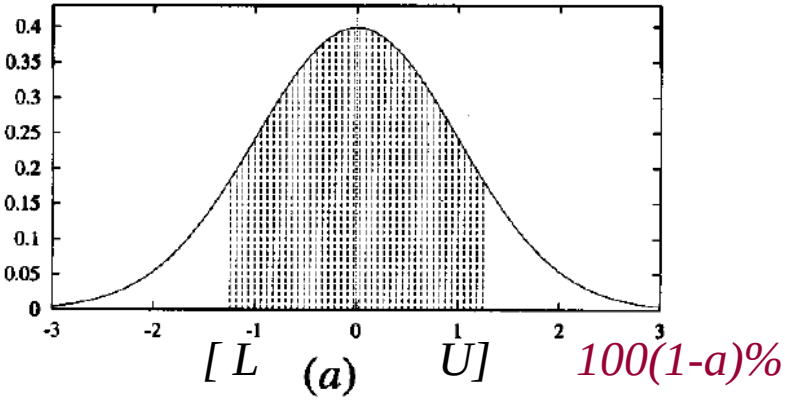
- از اینرو تقریب ما به صورت زیر خواهد بود:

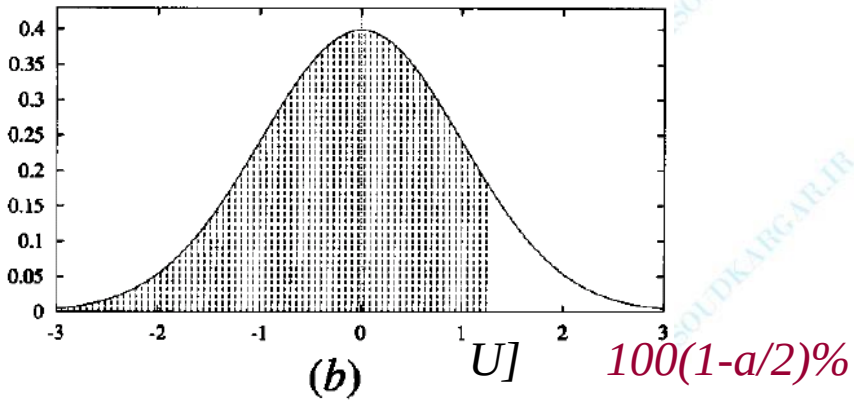
$$error_S(h) \pm Z_N \sqrt{\frac{error_S(h)(1-error_S(h))}{n}}$$

- برای بدست آوردن این رابطه دو تقریب زده شده است:
  1. در محاسبه انحراف معیار بجای $error_D(h)$ از $error_S(h)$ استفاده شده است.
  2. توزیع دوجمله‌ای با توزیع نرمال تقریب زده شده است.

حدود یکطرفه و دوطرفه

- فاصله بدست آمده در رابطه قبل یک فاصله دوطرفه است. گاهی لازم می‌شود که این فاصله به صورت یکطرفه بیان شود. بطور مثال:
- احتمال اینکه $error_D(h)$ حداکثر U باشد چقدر است؟
- با توجه به اینکه توزیع نرمال حول میانگین متقارن است، می‌توان یک فاصله اطمینان دوطرفه را به فاصله اطمینان یکطرفه معادل تبدیل نمود.





[L 100(1-a/2)%

اختلاف خطای فرضیه‌ها

- حالتی را در نظر بگیرید که دو فرضیه h_1, h_2 موجود باشند:
- h_1 بر روی مجموعه S_1 که شامل n_1 عضو است تست شده و h_2 بر روی مجموعه S_2 که شامل n_2 عضو بوده و دارای همان توزیع است تست گردیده است .

- می‌خواهیم بدانیم اختلاف خطای واقعی این دو فرضیه چیست؟

$$d \equiv error_D(h_1) - error_D(h_2)$$

تخمین‌زننده

- برای تخمین مقدار d از یک تخمین‌زننده استفاده می‌کنیم:

$$\hat{d} \equiv error_{S_1}(h_1) - error_{S_2}(h_2)$$

- نشان داده می‌شود که $\hat{d}$ تخمینی بدون بایاس از d را بدست می‌دهد یعنی:

$$E[\hat{d}] = d$$

انحراف معیار تخمین‌زننده

- از آنجائیکه برای مقادیر بزرگ نمونه توزیع احتمال $error_{S_1}(h_1)$ و $error_{S_2}(h_2)$ تقریبا نرمال است، لذا توزیع احتمال $\hat{d}$ را نیز می‌توان به صورت نرمال در نظر گرفت:

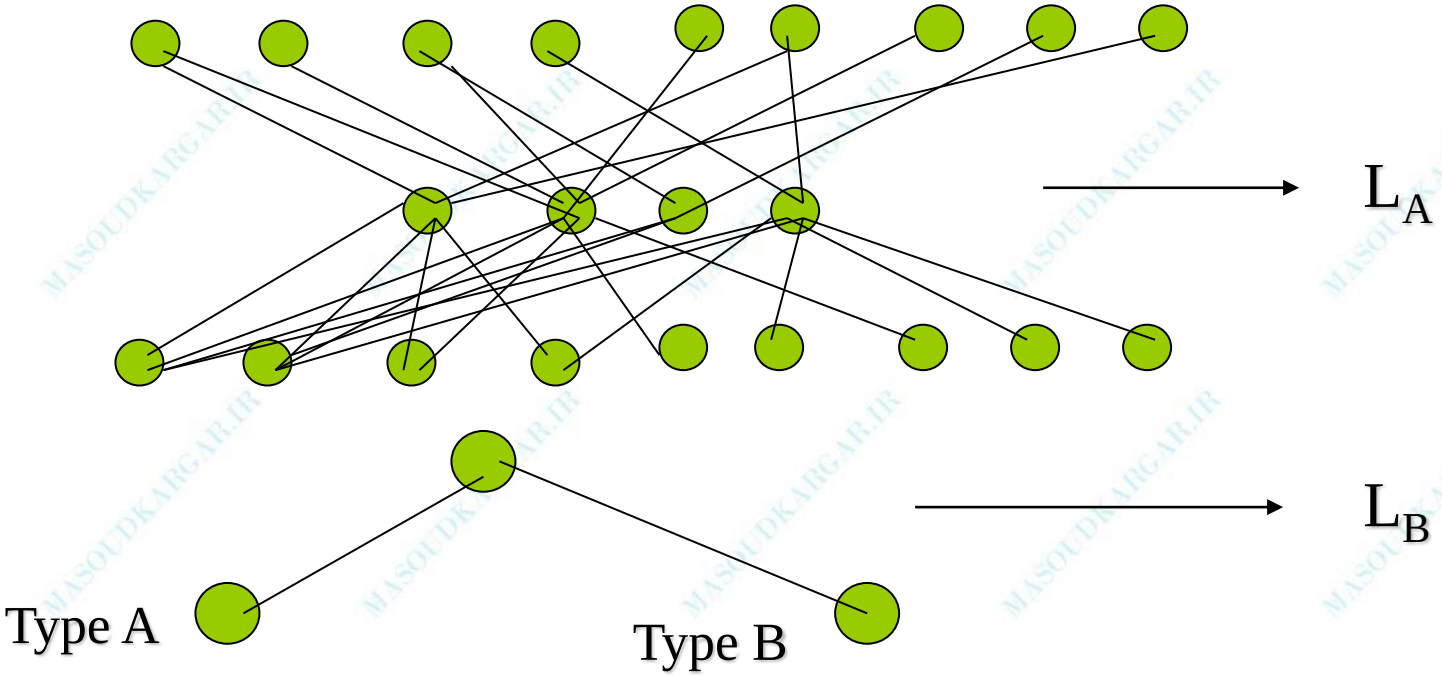
$$\sigma_{\hat{d}}^2 = \frac{error_{S_1}(h_1)(1-error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)(1-error_{S_2}(h_2))}{n_2}$$

- به همین ترتیب فاصله اطمینان این تقریب به صورت زیر خواهد بود.

$$\hat{d} \pm z_N \sqrt{\frac{error_{S_1}(h_1)(1-error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)(1-error_{S_2}(h_2))}{n_2}}$$

مقایسه الگوریتم‌های یادگیری

- چگونه می‌توان عملکرد دو الگوریتم یادگیری مختلف (مثل شبکه عصبی و درخت تصمیم) را مقایسه کرد؟



مقایسه الگوریتم‌های یادگیری

- روشهای مختلفی برای اینکار معرفی شده ولی هنوز روشی که بتواند اتفاق آرا را کسب کند ارائه نگردیده است!
- یک روش عبارت است از مقایسه میانگین عملکرد دو الگوریتم بر روی تمامی مجموعه‌های آموزشی با اندازه n که به صورت تصادفی از نمونه با توزیع D انتخاب می‌شوند.
- بعبارت دیگر می‌خواهیم مقدار اختلاف مورد انتظار درخطای آن دو را تخمین بزنیم.

$$E_{S \subset D}\left[error_D(L_A(S)) - error_D(L_B(S))\right]$$

داده کمی مشکل

- در عمل فقط تعداد کمی داده نمونه برای مقایسه دو الگوریتم وجود دارد. در چنین حالتی داده موجود به دو مجموعه داده آموزشی S_0 و مجموعه داده تست T_0 تقسیم می‌شود.از داده آموزشی برای آموزش هر دو الگوریتم استفاده شده و داده تست نیز برای ارزیابی هر دو الگوریتم استفاده می‌شود.
- در این صورت مقدار زیر برای مقایسه دو الگوریتم بکار می‌رود.

$$error_{T0}(L_A(S_0))\text{-}error_{T0}(L_B(S_0))$$

- ایراد این کار اینجاست که بجای استفاده از تمامی مجموعه‌های موجود در D فقط خطای موجود در مجموعه آموزشی مورد استفاده قرار می‌گیرد.

k-Fold Cross-Validation

- یک راه حل استفاده از الگوریتم زیر است:

1. Partition the available data D_0 into k disjoint subsets $T_1, T_2, \ldots, T_k$ of equal size, where this size is at least 30.
2. For i from 1 to k, do
   use T_i for the test set, and the remaining data for training set S_i
   - $S_i <= \{D_0 - T_i\}$
   - $h_A <= L_A(S_i)$
   - $h_B <= L_B(S_i)$
   - $\delta_i <= error_{Ti}(h_A) - error_{Ti}(h_B)$
3. Return the value $avg(\delta)$, where .

$$avg(\delta) = 1/k \sum_{i=1..k} \delta_i$$

اطمینان فاصله

- مقدار تقریبی فاصله اطمینان $N\%$ برای تخمین $E_{S \subset D0}[error_D(L_A(S))-error_D(L_B(S))]$ عبارت است از:

$$avg(\delta) \pm t_{N,k-1} s_{avg(\delta)}$$

- که در آن $t_{N,k-1}$ مقداری شبیه به Z_N بوده و مقادیر آن از جدول 5-6بدست می‌آید، $s_{avg(\delta)}$ تخمینی از انحراف معیار مربوط به توزیع $avg(\delta)$ می‌باشد:

$$s_{avg(\delta)} = \sqrt{1/k(k-1) \sum_{i=1}^{k} (\delta_i - avg(\delta))^2}$$

فاصله اطمینان

	Confidence level N			
	90%	95%	98%	99%
$\nu = 2$	2.92	4.30	6.96	9.92
$\nu = 5$	2.02	2.57	3.36	4.03
$\nu = 10$	1.81	2.23	2.76	3.17
$\nu = 20$	1.72	2.09	2.53	2.84
$\nu = 30$	1.70	2.04	2.46	2.75
$\nu = 120$	1.66	1.98	2.36	2.62
$\nu = \infty$	1.64	1.96	2.33	2.58

TABLE 5.6
Values of $t_{N,\nu}$ for two-sided confidence intervals.

As $\nu \rightarrow \infty$, $t_{N,\nu}$ approaches z_N.

- در $t_{N,\nu}$ به ν درجه آزادی گویند و برابر است با تعداد رخدادهای تصادفی مستقلی که باعث تولید متغیر تصادفی $avg(\delta)$ شده است. (تعداد مجموعه‌های افراز شده منهای یک)

Paired Test

- اگر تست دو فرضیه یادگیری با استفاده از مجموعه مثالهای یکسانی انجام شود *paired test* نامیده می‌شود.
- نتیجه چنین آزمایشاتی معمولا منجر به فواصل اطمینان بسته‌تری می‌گردد زیرا اختلاف مشاهده شده در خطا مربوط به اختلاف بین فرضیه‌هاست. در حالیکه وقتی فرضیه‌ها با استفاده از مجموعه داده‌های متفاوتی تست می‌شوند امکان تاثیرگذاری اختلاف بین دو مجموعه داده زیاد می‌شود.