

سوال ۱) write generalization error of an ensemble in terms of bias, variance and covariance

خطای تعمیم یک پیش‌بین منفرد (Single Predictor) می‌تواند به صورت جمعی از **Bias** و **Variance** بیان شود. در مورد یک سیستم تجمعی (Ensemble)، علاوه بر Bias و Variance، باید تأثیر همبستگی (Covariance) میان پیش‌بین‌ها نیز در نظر گرفته شود.

فرمول خطای تعمیم در سیستم تجمعی

خطای تعمیم برای یک سیستم تجمعی می‌تواند به صورت زیر بیان شود:

$$\text{Generalization Error} = \text{Bias}^2 + \frac{\text{Variance}}{M} + \text{Covariance} \cdot \frac{M-1}{M}$$

در این فرمول:

- $M$ : تعداد پیش‌بین‌ها در سیستم تجمعی.
- $\text{Bias}^2$ : سیستم که نشان‌دهنده خطای سیستماتیک است.
- $\frac{\text{Variance}}{M}$ : تأثیر پراکندگی پیش‌بین‌ها که با افزایش تعداد پیش‌بین‌ها کاهش می‌یابد.
- $\text{Covariance} \cdot \frac{M-1}{M}$ : تأثیر همبستگی بین پیش‌بین‌ها.

تحلیل فرمول:

- با افزایش تعداد پیش‌بین‌ها ( $M$ )، سهم واریانس کاهش می‌یابد ( $\frac{\text{Variance}}{M}$ ).
- اگر پیش‌بین‌ها مستقل باشند (همبستگی صفر)، قسمت همبستگی تأثیری ندارد.
- اگر همبستگی بین پیش‌بین‌ها زیاد باشد، خطای تعمیم به میزان قابل‌توجهی تحت تأثیر قرار می‌گیرد.

در نتیجه، برای کاهش خطای تعمیم در سیستم تجمعی، باید علاوه بر کاهش Bias و Variance، تلاش شود که پیش‌بین‌ها تا حد امکان کم‌همبسته باشند.

سوال ۲) explain why diversity between base classifiers of ensembles reduces overall generalization error of ensemble

تنوع (Diversity) بین طبقه‌بندهای پایه (Base Classifiers) در یک سیستم تجمعی (Ensemble) نقش بسیار مهمی در کاهش خطای تعمیم کلی دارد. دلیل این موضوع را می‌توان با توجه به نحوه عملکرد سیستم‌های تجمعی توضیح داد:

### ۱. جبران خطاها (Error Compensation)

طبقه‌بندهای مختلف، به دلیل تفاوت در ساختار، روش یادگیری یا داده‌هایی که روی آن‌ها آموزش دیده‌اند، معمولاً خطاهای مختلفی مرتکب می‌شوند. اگر این طبقه‌بندها تنوع داشته باشند:

- خطای یک طبقه‌بند توسط طبقه‌بند دیگر جبران می‌شود.
- در نتیجه، احتمال وقوع خطای مشابه بین همه طبقه‌بندها کاهش می‌یابد.
- این فرآیند باعث کاهش خطای کلی سیستم تجمعی می‌شود.

### ۲. کاهش واریانس (Variance Reduction)

در یک سیستم تجمعی، اگر طبقه‌بندها مستقل یا کم‌همبسته باشند (Diverse)، واریانس خطا در پیش‌بینی نهایی کاهش می‌یابد.

- طبقه‌بندهایی که به داده‌ها پاسخ‌های متفاوتی می‌دهند، اثر نوسانات ناشی از داده‌های نویزی یا متغیر را خنثی می‌کنند.
- این باعث می‌شود که سیستم تجمعی به داده‌های نویزی حساسیت کمتری داشته باشد و عملکرد بهتری در داده‌های جدید ارائه دهد.

### ۳. افزایش پوشش فضای داده‌ها:

تنوع باعث می‌شود که هر طبقه‌بند روی بخش‌های متفاوتی از فضای ویژگی‌ها تخصص داشته باشد. این موضوع باعث می‌شود که سیستم تجمعی توانایی پوشش دادن طیف گسترده‌ای از داده‌ها را پیدا کند و عملکرد کلی بهبود یابد.

### ۴. کاهش حساسیت به داده‌های نویزی:

در صورتی که داده‌ها دارای نویز باشند، طبقه‌بندهای متنوع احتمال کمتری دارند که همگی تحت تأثیر نویز قرار بگیرند. این موضوع به پایداری سیستم کمک می‌کند.

### مثال

فرض کنید دو طبقه‌بند داریم:

- طبقه‌بند اول در پیش‌بینی کلاس A خوب عمل می‌کند اما در پیش‌بینی کلاس B ضعیف است.
  - طبقه‌بند دوم در پیش‌بینی کلاس B خوب عمل می‌کند اما در کلاس A ضعیف است.
- اگر این دو طبقه‌بند را ترکیب کنیم (مثلاً با رأی‌گیری یا میانگین‌گیری)، سیستم تجمعی در هر دو کلاس عملکرد بهتری خواهد داشت.

## نتیجه‌گیری

تنوع بین طبقه‌بندهای پایه باعث می‌شود که:

- خطاهای پیش‌بین‌های منفرد کاهش یابد.
  - تأثیر نویز و همبستگی خطاها کم شود.
  - عملکرد کلی سیستم تجمعی در کاهش خطای تعمیم بهبود یابد.
- بنابراین، ایجاد تنوع بین طبقه‌بندهای پایه یکی از استراتژی‌های کلیدی در طراحی سیستم‌های تجمعی موفق است.