Задание 3. Поляков Даниил, совместно с 494 гр.

1. Покажите справедливость bias-variance decomposition:

$$\mathbf{E}_{x,y}\mathbf{E}_{X^\ell}(y - a_{X^\ell}(x))^2 = \mathbf{E}_{x,y}(y - \mathbf{E}(\mathbf{y}|\mathbf{x}))^2 + \mathbf{E}_{x,y}(\mathbf{E}(\mathbf{y}|\mathbf{x}) -$$

$$- \mathbf{E}_{X^\ell} a_{X^\ell}(x))^2 + \mathbf{E}_{x,y}\mathbf{E}_{X^\ell}(a_{X^\ell}(x) - \mathbf{E}_{X^\ell} a_{X^\ell}(x))^2$$

По свойствам условного мат.ожидания:

$$\mathbf{E}_{x,y}(y - a_{X^\ell}(x))^2 = \mathbf{E}_{x,y}(\mathbf{E}_{x,y}(y - a_{X^\ell}(x))^2|x))$$

Мы знаем, что в данных есть некоторый шум: $y = f(x) + \epsilon$. Тогда

$$\mathbf{E}_{x,y}(y - a_{X^\ell}(x))^2|x) = \mathbf{E}_{x,y}(a_{X^\ell}(x) - f(x) - \epsilon)^2|x) =$$

$$= \mathbf{E}_{x,y}(a_{X^\ell}(x) - f(x))^2 - 2\epsilon(a(x) - f(x)) + \epsilon^2|x) =$$

с учетом того, что ε - случайный шум, получим

$$= \mathbf{E}_{x,y}(a_{X^\ell}(x) - f(x))^2|x) - 2\mathbf{E}_{x,y}(\varepsilon|x)\mathbf{E}_{x,y}(a_{X^\ell}(x) - f(x))|x) + \mathbf{E}_{x,y}(\varepsilon^2|x)$$

$$= \mathbf{E}_{x,y}(a_{X^\ell}(x) - f(x))^2|x) - 2\mathbf{E}_{x,y}(\varepsilon)\mathbf{E}_{x,y}(a_{X^\ell}(x) - f(x))|x) + \mathbf{E}_{x,y}(\varepsilon^2)$$

$$= \mathbf{E}_{x,y}(a_{X^\ell}(x) - f(x))^2|x) - 2 \cdot 0 \cdot \mathbf{E}_{x,y}(a_{X^\ell}(x) - f(x))|x) + \sigma^2$$

$$= \mathbf{E}_{x,y}(a_{X^\ell}(x) - f(x))^2|x) + \sigma^2$$

Обозначим $\mathbf{E}_{x,y}(a(x)) := \overline{a}(x)$. Тогда

$$\mathbf{E}_{x,y}((a_{X^\ell}(x) - \overline{a}_{X^\ell}(x)) + (\overline{a}(x) - f(x))^2|x) =$$

$$\mathbf{E}_{x,y}((a_{X^\ell}(x) - \overline{a}(x)))^2 - 2\mathbf{E}((a_{X^\ell}(x) - \overline{a}(x)))((\overline{a}(x) - f(x))|x) + \mathbf{E}((\overline{a}(x) - f(x))^2|x) =$$

так как $(\overline{a}(x) - f(x)) -$ просто число, то эту скобку можно вынести за знак мат.ожидания:

$$= \mathbf{E}_{x,y}((a_{X^\ell}(x) - \overline{a}(x)))^2 - 2(\overline{a}(x) - f(x)|x)\mathbf{E}((a_{X^\ell}(x) - \overline{a}(x))) + \mathbf{E}((\overline{a}(x) - f(x))^2|x) =$$

$$= \mathbf{E}_{x,y}((a_{X^\ell}(x) - \overline{a}(x)))^2 - 2(\overline{a}(x) - f(x)|x)\mathbf{E}((\overline{a}(x) - \overline{a}(x))) + \mathbf{E}((\overline{a}(x) - f(x))^2|x) =$$

$$= \mathbf{E}_{x,y}((a_{X^\ell}(x) - \overline{a}(x)))^2 + (\overline{a}(x) - f(x))^2$$

Учитывая, что $\sigma^2 = \mathbf{E}_{x,y}(y - \mathbf{E}(\mathbf{y}|\mathbf{x}))^2$, окончательно получим:

$$\mathbf{E}_{x,y}(y - a_{X^\ell}(x))^2 = \mathbf{E}_{x,y}((a_{X^\ell}(x) - \overline{a}(x)))^2 + (\overline{a}(x) - f(x))^2 + \mathbf{E}_{x,y}(y - \mathbf{E}(\mathbf{y}|\mathbf{x}))^2 =$$

$$= Variance(x) + (Bias(x))^2 + noise^2$$

2.2 Корреляция ответов базовых алгоритмов

Показать, что если есть M одинаково распределенных величин с дисперсией σ^2, любые две из которых имеют положительную корреляцию ρ, то

$$\mathbb{D}\overline{xi} = \rho\sigma^2 + (1-\rho)\frac{\sigma^2}{M}$$

$$\mathbb{D}\overline{\xi_i} = \mathbb{D}\left(\frac{1}{M}\sum_{i=1}^{M}\xi_i\right) = \frac{1}{M^2}\mathbb{D}\left(\sum_{i=1}^{M}\xi_i\right)$$

Учмтывая, что $\mathbb{D}(x+y) = \mathbb{D}x + \mathbb{D}y + 2\text{cov}(x;y)$, получим

$$= \frac{1}{M^2}\left(\sum_{i=1}^{M}\mathbb{D}\xi_i \quad + \sum_{i=1,j=1,i\neq j}^{M}\text{cov}(\xi_i;\xi_j)\right) =$$

По определению корреляции двух случайных величин:

$$\text{corr}(\xi_1,\xi_2) = \frac{\text{cov}(\xi_1,\xi_2)}{\sigma_{\xi_1}\sigma_{\xi_2}} \Rightarrow \text{cov}(\xi_1,\xi_2) = \sigma^2\cdot\text{corr}(\xi_1,\xi_2) = \sigma^2\cdot\rho$$

Продолжим цепочку равенств

$$= \frac{1}{M^2}\left(M\sigma^2 + (M^2-M)\cdot\sigma^2\rho\right) = \frac{\sigma^2}{M} + \left(1-\frac{1}{M}\right)\sigma^2\rho = \sigma^2\rho + \frac{\sigma^2}{M}(1-\rho)$$

что и требовалось показать.

2.3 Смещение и разброс в бэггинге

Пусть ответы всех базовых алгоритмов распределены одинаково. Выясните как соотносятся смещение и разброс для композиции с теми же параметрами для базовых алгоритмов, если композиция строится с помощью бэггинга:

$$a(x) = \frac{1}{M}\sum_{i=1}^{M}a_i(x)$$

Как было показано ранее

$$Variance(x) = \mathbf{E}_{X,Y,x,y}((a(x)-\overline{a}(x))^2), \qquad Bias = \overline{a}(x) - f(x)$$

$$\mathbf{E}_{X,Y,x,y}(a(x)) = \frac{1}{M}\mathbf{E}_{X,Y,x,y}\sum_{i=1}^{M}(a_i(x)) = \mathbf{E}_{X,Y,x,y}a_1(x)$$

Таким образом мат.ожидание композиции не изменилось, следовательно не изменился и $bias$.

Выясним, какой вид у $Variance$ (дисперсии). Пусть любая пара алгоритмов имеет положительную корреляцию ρ. Тогда как показано в предыдущем номере

$$\mathbf{Var}(a(x)) = \mathbf{Var}(a_1(x)) \left(\frac{1}{M} + \left(1 - \frac{1}{M} \right) \rho \right)$$

Таким образом видим, что разброс (дисперсия) композиции алгоритмов тем ниже, чем менее скоррелированы обученные базовые алгоритмы. В этом и состоит идея обучения базовых алгоритмов на случайных подвыборках (бэггинг) и на слуайном множестве признаков.