CSC 2515: Introduction to Machine Learning

Lecture 4: Bias-Variance Decomposition, Ensemble Method I: Bagging

Rahul G. Krishnan[1]

University of Toronto and Vector Institute

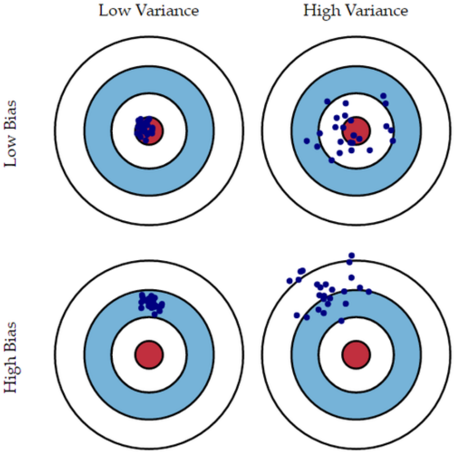
[1] Credit for slides goes to many members of the ML Group at the U of T, and beyond, including (recent past): Amir-Massoud Farahmand, Roger Grosse, Murat Erdogdu, Richard Zemel, Juan Felipe Carrasquilla, Emad Andrews, and myself.

Table of Contents

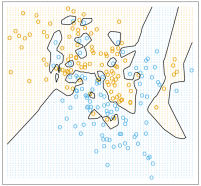
Today

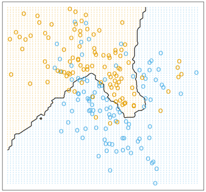
- Closer look at what determines the error of ML algorithm
- Bootstrap Aggregation (Bagging)
- Skills to Learn
  - Understanding the bias-variance decomposition
  - The concept behind Bagging and why it works
  - Random Forests

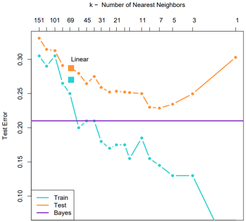
Bias-Variance Decomposition



- Recall that overly simple models underfit the data, and overly complex models overfit.







- We quantify this effect in terms of the bias-variance decomposition.

Bias-Variance Decomposition for the Mean Estimator

- For the next few slides, we consider the simple problem of estimating the mean of a random variable (r.v.) using data.
- Consider a r.v. Y with an unknown distribution p. This random variable has an (unknown) mean $m = \mathbb{E}[Y]$ and variance $\sigma^2 = \mathrm{Var}[Y] = \mathbb{E}\left[(Y - m)^2\right]$.
- Given: a dataset $\mathcal{D} = \{Y_1, \ldots, Y_n\}$ with independently sampled $Y_i \sim p$.
- How can we estimate m using $\mathcal{D}$?

Bias-Variance Decomposition for the Mean Estimator

- Given: a dataset $\mathcal{D} = \{Y_1, \ldots, Y_n\}$ with independently sampled $Y_i \sim p$.
- Consider an algorithm that receives $\mathcal{D}$, does some processing on data, and outputs a number. The goal of this algorithm is to provide an estimate of m. Let us denote it by $h(\mathcal{D})$.
- Some good and bad examples:
  - Sample average: $h(\mathcal{D}) = \frac{1}{n} \sum_{i=1}^{n} Y_i$
  - Single-sample estimator: $h(\mathcal{D}) = Y_1$
  - Zero estimator: $h(\mathcal{D}) = 0$
- How well do they perform?

Bias-Variance Decomposition for the Mean Estimator

- How can we assess the performance of a particular $h(\mathcal{D})$?
- Ideally, we want $h(\mathcal{D})$ to be exactly equal to $m = \mathbb{E}[Y]$. But this might be too much to ask. (Q: Why?)
- What we can hope for is that $h(\mathcal{D}) \approx m$.
- How can we quantify the accuracy of approximation?

Bias-Variance Decomposition for the Mean Estimator

- We use the squared error $\text{err}(\mathcal{D}) = |h(\mathcal{D}) - m|^2$ as a measure of quality. This is the familiar squared error loss function in regression.
- The error $\text{err}(\mathcal{D})$ is a r.v. itself. (Q: Why?)
- For a dataset $\mathcal{D} = \{Y_1, \ldots, Y_n\}$ the loss $\text{err}(D)$ might be small, but for another $\mathcal{D}' = \{Y_1', \ldots, Y_n'\}$ (still with $Y_i' \sim p$) the loss $\text{err}(D')$ might be large. We would like to quantify the average error.
- We focus on the expectation of $\text{err}(\mathcal{D})$, i.e.,

$$\mathbb{E}\left[\text{err}(\mathcal{D})\right] = \mathbb{E}_{\mathcal{D}}\left[|h(\mathcal{D}) - m|^2\right].$$

- Note that the dataset $\mathcal{D}$ is random and this expectation is w.r.t. its randomness.

Bias-Variance Decomposition for the Mean Estimator

- We would like to understand what determines $\mathbb{E}_{\mathcal{D}}\left[|h(\mathcal{D})-m|^2\right]$ by looking more closely at it.
- We can decompose $\mathbb{E}_{\mathcal{D}}\left[|h(\mathcal{D})-m|^2\right]$ by adding and subtracting $\mathbb{E}_{\mathcal{D}}[h(\mathcal{D})]$ inside $|\cdot|$ and expanding:

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[|h(\mathcal{D})-m|^2\right] =& \mathbb{E}_{\mathcal{D}}\left[|h(\mathcal{D})-\mathbb{E}_{\mathcal{D}}[h(\mathcal{D})]+\mathbb{E}_{\mathcal{D}}[h(\mathcal{D})]-m|^2\right] \\ =& \mathbb{E}_{\mathcal{D}}\left[|h(\mathcal{D})-\mathbb{E}_{\mathcal{D}}[h(\mathcal{D})]|^2\right]+\mathbb{E}_{\mathcal{D}}\left[|\mathbb{E}_{\mathcal{D}}[h(\mathcal{D})]-m|^2\right]+ \\ & 2\mathbb{E}_{\mathcal{D}}\left[(h(\mathcal{D})-\mathbb{E}_{\mathcal{D}}[h(\mathcal{D})])(\mathbb{E}_{\mathcal{D}}[h(\mathcal{D})]-m)\right].\end{aligned}$$

- What is the intuition of term $\mathbb{E}_{\mathcal{D}}[h(\mathcal{D})]$?
- Let us simplify the right hand side (RHS).

Bias-Variance Decomposition for the Mean Estimator

$$\mathbb{E}_{\mathcal{D}}\left[|h(\mathcal{D}) - m|^2\right] = \mathbb{E}_{\mathcal{D}}\left[|h(\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[h(\mathcal{D})]|^2\right] + \mathbb{E}_{\mathcal{D}}\left[|\mathbb{E}_{\mathcal{D}}[h(\mathcal{D})] - m|^2\right] + 2\mathbb{E}_{\mathcal{D}}\left[(h(\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[h(\mathcal{D})])(\mathbb{E}_{\mathcal{D}}[h(\mathcal{D})] - m)\right].$$

- Recall that if X is a random variable and f is a function, the quantity $f(X)$ is a random variable. But its expectation $\mathbb{E}[f(X)]$ is not. We can say that the expectation takes the randomness away. So $\mathbb{E}_{\mathcal{D}}[h(\mathcal{D})]$ is not a random variable anymore.
- For the second term, we have

$$\mathbb{E}_{\mathcal{D}}\left[|\mathbb{E}_{\mathcal{D}}[h(\mathcal{D})] - m|^2\right] = |\mathbb{E}_{\mathcal{D}}[h(\mathcal{D})] - m|^2 .$$

- Q: Suppose that $\mathcal{D} = \{Y_1, \ldots, Y_n\}$ with $Y_i \sim p$ with mean m and variance σ^2. Define $h(\mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} Y_i^2$. What is $\mathbb{E}_{\mathcal{D}}[h(\mathcal{D})]$?

Bias-Variance Decomposition for the Mean Estimator

$$\mathbb{E}_{\mathcal{D}}\left[|h(\mathcal{D})-m|^2\right] = \mathbb{E}_{\mathcal{D}}\left[|h(\mathcal{D})-\mathbb{E}_{\mathcal{D}}[h(\mathcal{D})]|^2\right] + \mathbb{E}_{\mathcal{D}}\left[|\mathbb{E}_{\mathcal{D}}[h(\mathcal{D})]-m|^2\right] + 2\mathbb{E}_{\mathcal{D}}\left[(h(\mathcal{D})-\mathbb{E}_{\mathcal{D}}[h(\mathcal{D})])(\mathbb{E}_{\mathcal{D}}[h(\mathcal{D})]-m)\right].$$

- Let us consider $\mathbb{E}_{\mathcal{D}}\left[(h(\mathcal{D})-\mathbb{E}_{\mathcal{D}}[h(\mathcal{D})])(\mathbb{E}_{\mathcal{D}}[h(\mathcal{D})]-m)\right]$.
- To reduce the clutter, we denote $\bar{m} = \mathbb{E}_{\mathcal{D}}[h(\mathcal{D})]$, i.e., the expected value of the estimator.
- Note that $\bar{m}$ is an expectation of a r.v., so it is not random. This means that $\mathbb{E}[\bar{m}h(\mathcal{D})] = \bar{m}\mathbb{E}[h(\mathcal{D})]$.
- We have

$$\mathbb{E}_{\mathcal{D}}\left[(h(\mathcal{D})-\mathbb{E}_{\mathcal{D}}[h(\mathcal{D})])(\mathbb{E}_{\mathcal{D}}[h(\mathcal{D})]-m)\right] =$$
$$\mathbb{E}_{\mathcal{D}}[(h(\mathcal{D})-\bar{m})(\bar{m}-m)] = (\bar{m}-m)\underbrace{(\mathbb{E}_{\mathcal{D}}[h(\mathcal{D})]-\bar{m})}_{=0} = 0$$

The third term is zero.

Bias-Variance Decomposition for the Mean Estimator

Bias-Variance Decomposition

$$\mathbb{E}_{\mathcal{D}}\left[\left|h(\mathcal{D}) - m\right|^2\right] = \underbrace{\left|\mathbb{E}_{\mathcal{D}}\left[h(\mathcal{D})\right] - m\right|^2}_{\text{bias}} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left|h(\mathcal{D}) - \mathbb{E}_{\mathcal{D}}\left[h(\mathcal{D})\right]\right|^2\right]}_{\text{variance}}.$$

- Bias: The error of the expected estimator (over draws of dataset $\mathcal{D}$) compared to the mean $m = \mathbb{E}\left[Y\right]$ of the random variable Y.
- Variance: The variance of a single estimator $h(\mathcal{D})$ (whose randomness comes from $\mathcal{D}$).
- This is for an estimator of a mean of a random variable. We shall extend this decomposition to more general estimators too.

Bias-Variance Decomposition for the Mean Estimator: Examples

Bias-Variance Decomposition

$$\mathbb{E}_{\mathcal{D}}\left[|h(\mathcal{D}) - m|^2\right] = \underbrace{|\mathbb{E}_{\mathcal{D}}[h(\mathcal{D})] - m|^2}_{\text{bias}} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[|h(\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[h(\mathcal{D})]|^2\right]}_{\text{variance}}.$$

- Let us compute the bias and variance of a few estimators. Recall that $m = \mathbb{E}[Y]$ and $\sigma^2 = \mathrm{Var}\{Y\} = \mathbb{E}\left[(Y - m)^2\right]$.
- Sample average: $h(\mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} Y_i$.
  - Bias $|\mathbb{E}_{\mathcal{D}}[h(\mathcal{D})] - m|^2 = |\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] - m|^2 = |\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y_i] - m|^2 = |\frac{1}{n}\sum_{i=1}^{n} m - m|^2 = 0.$
  - Variance: $\mathbb{E}\left[|h(\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[h(\mathcal{D})]|^2\right] = \mathbb{E}\left[|\frac{1}{n}\sum_{i=1}^{n} Y_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right]|^2\right] =$ $\mathbb{E}\left[|\frac{1}{n}\sum_{i=1}^{n}(Y_i - m)|^2\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[(Y_i - m)^2\right] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.$
  - $\mathbb{E}_{\mathcal{D}}\left[|h(\mathcal{D}) - m|^2\right] = \text{bias} + \text{variance} = 0 + \frac{\sigma^2}{n}.$

Bias-Variance Decomposition for the Mean Estimator: Examples

Bias-Variance Decomposition

$$\mathbb{E}_{\mathcal{D}}\left[|h(\mathcal{D}) - m|^2\right] = \underbrace{|\mathbb{E}_{\mathcal{D}}[h(\mathcal{D})] - m|^2}_{\text{bias}} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[|h(\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[h(\mathcal{D})]|^2\right]}_{\text{variance}}.$$

- Single-sample estimator: $h(\mathcal{D}) = Y_1$
  - The algorithm behind this estimator only looks at the first data point and ignores the rest.
  - Bias $|\mathbb{E}_{\mathcal{D}}[h(\mathcal{D})] - m|^2 = |\mathbb{E}[Y_1] - m|^2 = |m - m|^2 = 0$.
  - Variance: $\mathbb{E}\left[|h(\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[h(\mathcal{D})]|^2\right] = \mathbb{E}\left[|Y_1 - \mathbb{E}[Y_1]|^2\right] = \sigma^2$.
  - $\mathbb{E}_{\mathcal{D}}\left[|h(\mathcal{D}) - m|^2\right] = \text{bias} + \text{variance} = 0 + \sigma^2$.

Bias-Variance Decomposition for the Mean Estimator: Examples

Bias-Variance Decomposition

$$\mathbb{E}_{\mathcal{D}}\left[|h(\mathcal{D}) - m|^2\right] = \underbrace{|\mathbb{E}_{\mathcal{D}}\left[h(\mathcal{D})\right] - m|^2}_{\text{bias}} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[|h(\mathcal{D}) - \mathbb{E}_{\mathcal{D}}\left[h(\mathcal{D})\right]|^2\right]}_{\text{variance}}.$$

- Zero estimator: $h(\mathcal{D}) = 0$
  - The algorithm behind this estimator does not look at data and always outputs zero. (We do not really want to use it in practice.)
  - Bias $|\mathbb{E}_{\mathcal{D}}\left[h(\mathcal{D})\right] - m|^2 = |0 - m|^2 = m^2$.
  - Variance: $\mathbb{E}\left[|h(\mathcal{D}) - \mathbb{E}_{\mathcal{D}}\left[h(\mathcal{D})\right]|^2\right] = \mathbb{E}\left[|0 - \mathbb{E}\left[0\right]|^2\right] = 0$.
  - $\mathbb{E}_{\mathcal{D}}\left[|h(\mathcal{D}) - m|^2\right] = \text{bias} + \text{variance} = m^2 + 0$.

Bias-Variance Decomposition for the Mean Estimator: Examples

- Summary:
  - Sample average: $\mathbb{E}_{\mathcal{D}}\left[|h(\mathcal{D}) - m|^2\right] = \text{bias} + \text{variance} = 0 + \frac{\sigma^2}{n}$
  - Single-sample estimator: $\mathbb{E}_{\mathcal{D}}\left[|h(\mathcal{D}) - m|^2\right] = \text{bias} + \text{variance} = 0 + \sigma^2$.
  - Zero estimator: $\mathbb{E}_{\mathcal{D}}\left[|h(\mathcal{D}) - m|^2\right] = \text{bias} + \text{variance} = m^2 + 0$.
- These estimators show different behaviour of bias and variance.
  - The zero estimator has no variance (surprising?), but potentially a lot of bias (unless we are "lucky" and m is in fact very close to 0).
  - The sample average has zero bias, but in general it has a non-zero variance.
    - Q: When does it have a zero variance?

Bias-Variance Decomposition for the Mean Estimator

- We could also define error as

$$\mathbb{E}_{\mathcal{D},Y}\left[|h(\mathcal{D}) - Y|^2\right]$$

instead of $\mathbb{E}_{\mathcal{D}}\left[|h(\mathcal{D}) - m|^2\right]$. This measures the expected squared error of $h(\mathcal{D})$ compared to Y instead of the mean $m = \mathbb{E}[Y]$.

- We have a similar decomposition:

$$\begin{aligned}
\mathbb{E}\left[|h(\mathcal{D}) - Y|^2\right] =& \mathbb{E}\left[|h(\mathcal{D}) - m + m - Y|^2\right] \\
=& \mathbb{E}\left[|h(\mathcal{D}) - m|^2\right] + \mathbb{E}\left[|m - Y|^2\right] + \\
& 2\mathbb{E}\left[(h(\mathcal{D}) - m)(m - Y)\right].
\end{aligned}$$

- The last term is zero because

$$\begin{aligned}
\mathbb{E}\left[(h(\mathcal{D}) - m)(m - Y)\right] &= \mathbb{E}\left[\mathbb{E}\left[(h(\mathcal{D}) - m)(m - Y) \mid \mathcal{D}\right]\right] \\
&= \mathbb{E}\left[(h(\mathcal{D}) - m)\,\mathbb{E}\left[m - Y \mid \mathcal{D}\right]\right] = 0.
\end{aligned}$$

Bias-Variance Decomposition for the Mean Estimator

Bias-Variance Decomposition

$$\mathbb{E}\left[|h(\mathcal{D}) - Y|^2\right] = \underbrace{|\mathbb{E}_{\mathcal{D}}\left[h(\mathcal{D})\right] - m|^2}_{\text{bias}} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[|h(\mathcal{D}) - \mathbb{E}_{\mathcal{D}}\left[h(\mathcal{D})\right]|^2\right]}_{\text{variance}} + \underbrace{\mathbb{E}\left[|Y - m|^2\right]}_{\text{Bayes error}}.$$

- We have an additional term of $\mathbb{E}\left[|m - Y|^2\right] = \sigma^2$. This is the variance of Y. This comes from the randomness of the r.v. Y and cannot be avoided. This is called the Bayes error.

Bias-Variance Decomposition for the General Case

- What about the bias-variance decomposition for a machine learning algorithm such as a regression estimator or a classifier?
- Two importance issues to be addressed:
  - We are not trying to estimate a single real-valued number ($h(\mathcal{D}) \in \mathbb{R}$) anymore, but a function over input $\mathbf{x}$. How can we measure the error in this case?
  - When we only wanted to estimate the mean, the “best” solution was $m = \mathbb{E}[Y]$. What is the best solution here?

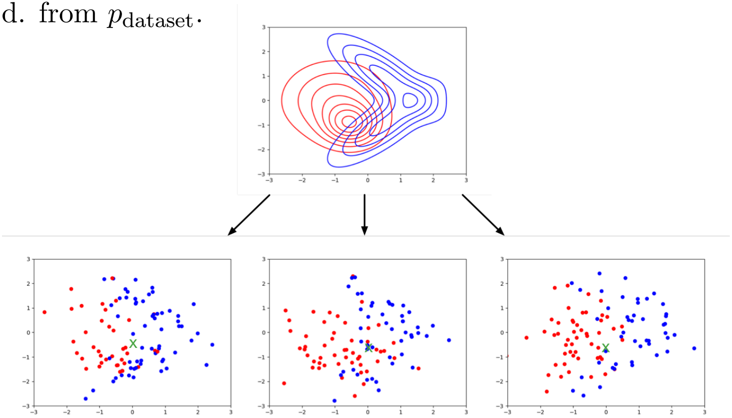
Bias-Variance Decomposition for the General Case

- Suppose that the training set $\mathcal{D}$ consists of N pairs $(\mathbf{x}^{(i)}, t^{(i)})$ sampled independent and identically distributed (i.i.d.) from a sample generating distribution p_{sample}, i.e., $(\mathbf{x}^{(i)}, t^{(i)}) \sim p_{\text{sample}}$.
- We consider the marginal distributions $p_{\mathbf{x}}$ and the distribution of t conditioned on $\mathbf{x}$ by $p(t|\mathbf{x})$:
  - $p_{\mathbf{x}}(\mathbf{x}) = \int p_{\text{sample}}(\mathbf{x}, t)\mathrm{d}t$
  - $p(t|\mathbf{x}) = \frac{p_{\text{sample}}(\mathbf{x},t)}{p_{\mathbf{x}}(\mathbf{x})}$
- Let p_{dataset} denote the induced distribution over training sets, i.e. $\mathcal{D} \sim p_{\text{dataset}}$.
  - We have that

$$p_{\text{dataset}}\left((\mathbf{x}^{(1)}, t^{(1)}), \ldots, (\mathbf{x}^{(N)}, t^{(N)})\right) = \prod_{i=1}^{N} p_{\text{sample}}((\mathbf{x}^{(i)}, t^{(i)})).$$

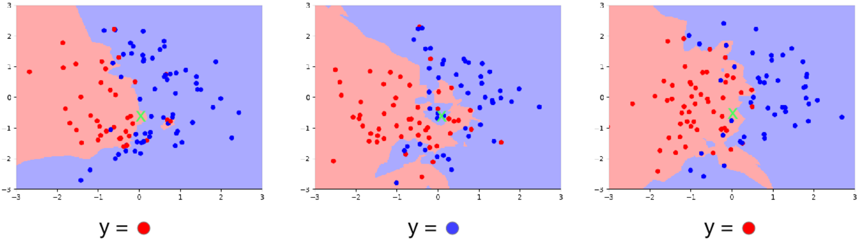
Bias-Variance Decomposition for the General Case

- Pick a fixed query point $\mathbf{x}$ (denoted with a green $\times$).
- Consider an experiment where we sample lots of training datasets i.i.d. from p_{dataset}.



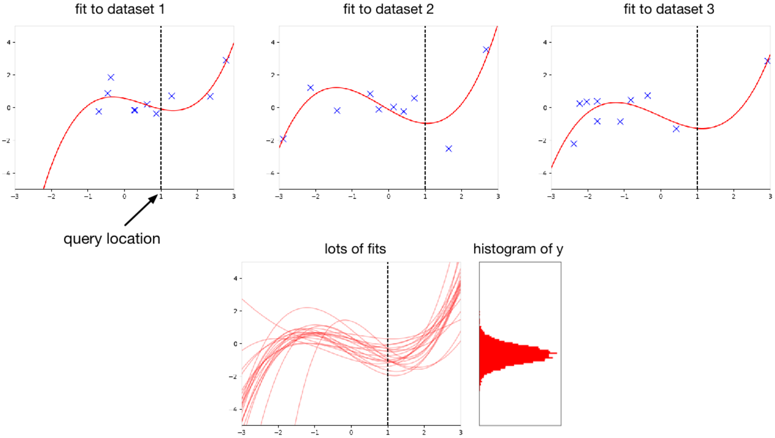
Bias-Variance Decomposition for the General Case

- Let us run our learning algorithm on each training set $\mathcal{D}$, producing a regressor or classifier $h(\mathcal{D}) : \mathcal{X} \to \mathcal{T}$.
- As $\mathcal{D}$ is random, and $h(\mathcal{D})$ is a function of $\mathcal{D}$, the function $h(\mathcal{D})$ is a random function.
- Fix a query point $\mathbf{x}$. We use $h(\mathcal{D})$ to predict the output at $\mathbf{x}$, i.e., $y = h(\mathbf{x}; \mathcal{D})$.
- y is a random variable, where the randomness comes from the choice of training set
  - $\mathcal{D}$ is random $\implies$ $h(\cdot; \mathcal{D})$ is random $\implies$ $h(\mathbf{x}; \mathcal{D})$ is random



Bias-Variance Decomposition for the General Case

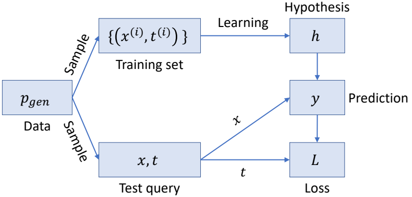
Here is the analogous setup for regression:



Since $y = h(\mathbf{x}; \mathcal{D})$ is a random variable, we can talk about its expectation, variance, etc. over the distribution of training sets p_{dataset}

Bias-Variance Decomposition for the General Case

- Recap of the setup:



- When $\mathbf{x}$ is fixed, this is very similar to the mean estimator case.
  - Recall that we had $\mathbb{E}_{\mathcal{D}}\left[|h(\mathcal{D}) - m|^2\right]$. In the mean estimator, $h(\mathcal{D})$ was a scalar r.v., but here we have $h(\mathcal{D}) : \mathcal{X} \to \mathcal{T}$.
- Can we have a bias-variance decomposition for a $h(\mathcal{D}) : \mathcal{X} \to \mathcal{T}$?
- Two questions:
  - What should replace m in the error decomposition?
  - How should we evaluate the performance when $\mathbf{x}$ is random?

Bayes Optimal Prediction

Proposition: For a fixed $\mathbf{x}$, the best estimator is the conditional expectation of the target value $y_*(\mathbf{x}) = \mathbb{E}\left[t|\mathbf{x}\right]$ (Distribution of $t \sim p(t|\mathbf{x})$), i.e.,

$$y_*(\mathbf{x}) = \underset{y}{\operatorname{argmin}}\, \mathbb{E}[(y-t)^2 \,|\, \mathbf{x}].$$

- **Proof:** Start by conditioning on (a fixed) $\mathbf{x}$. For any fixed y, we have

$$\begin{aligned}
\mathbb{E}[(y-t)^2 \,|\, \mathbf{x}] &= \mathbb{E}[y^2 - 2yt + t^2 \,|\, \mathbf{x}] \\
&= y^2 - 2y\mathbb{E}[t \,|\, \mathbf{x}] + \mathbb{E}[t^2 \,|\, \mathbf{x}] \\
&= y^2 - 2y\mathbb{E}[t \,|\, \mathbf{x}] + \mathbb{E}[t \,|\, \mathbf{x}]^2 + \mathrm{Var}[t \,|\, \mathbf{x}] \\
&= y^2 - 2yy_*(\mathbf{x}) + y_*(\mathbf{x})^2 + \mathrm{Var}[t \,|\, \mathbf{x}] \\
&= (y - y_*(\mathbf{x}))^2 + \mathrm{Var}[t \,|\, \mathbf{x}].
\end{aligned}$$

- The first term is nonnegative, and can be made 0 by setting $y = y_*(\mathbf{x})$.
- The second term does not depend on y. It corresponds to the inherent unpredictability, or noise, of the targets, and is called the Bayes error or irreducible error.
  - This is the best we can ever hope to do with any learning algorithm. An algorithm that achieves it is Bayes optimal.

Bias-Variance Decomposition for the General Case

- For each query point $\mathbf{x}$, the expected loss is different. We are interested in quantifying how well our estimator performs over the distribution p_{sample}. That is, the error measure is

$$\begin{aligned} \text{err}(\mathcal{D}) &= \mathbb{E}_{\mathbf{x}\sim p_{\mathbf{x}}}\left[\left|h(\mathbf{x};\mathcal{D}) - y_*(\mathbf{x})\right|^2\right] \\ &= \int \left|h(\mathbf{x};\mathcal{D}) - y_*(\mathbf{x})\right|^2 p_{\mathbf{x}}(\mathbf{x})\mathrm{d}\mathbf{x}. \end{aligned}$$

- This is similar to $\text{err}(\mathcal{D}) = |h(\mathcal{D}) - m|^2$ of the Mean Estimator case, except that
  - The ideal estimator is $y_*(\mathbf{x})$ and not m.
  - We take average over $\mathbf{x}$ according to the probability distribution $p_{\mathbf{x}}$.
- As before, $\text{err}(\mathcal{D})$ is random due to the randomness of $\mathcal{D} \sim p_{\text{dataset}}$.
- We focus on the expectation of $\text{err}(\mathcal{D})$, i.e.,

$$\mathbb{E}\left[\text{err}(\mathcal{D})\right] = \mathbb{E}_{\mathcal{D}\sim p_{\text{dataset}},\mathbf{x}\sim p_{\mathbf{x}}}\left[\left|h(\mathbf{x};\mathcal{D}) - y_*(\mathbf{x})\right|^2\right].$$

Bias-Variance Decomposition for the General Case

- To obtain the bias-variance decomposition of

$$\mathbb{E}\left[\mathrm{err}(\mathcal{D})\right] = \mathbb{E}_{\mathcal{D}\sim p_{\text{dataset}},\mathbf{x}\sim p_{\mathbf{x}}}\left[\left|h(\mathbf{x};\mathcal{D}) - y_*(\mathbf{x})\right|^2\right],$$

we add and subtract $\mathbb{E}_{\mathcal{D}}\left[h(\mathbf{x};\mathcal{D}) \mid \mathbf{x}\right]$ inside $|\cdot|$ (similar to before):

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D},\mathbf{x}}\left[\left|h(\mathbf{x};\mathcal{D}) - y_*(\mathbf{x})\right|^2\right] = \\
&\mathbb{E}_{\mathcal{D},\mathbf{x}}\left[\left|h(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}\left[h(\mathbf{x};\mathcal{D}) \mid \mathbf{x}\right] + \mathbb{E}_{\mathcal{D}}\left[h(\mathbf{x};\mathcal{D}) \mid \mathbf{x}\right] - y_*(\mathbf{x})\right|^2\right] = \\
&\mathbb{E}_{\mathcal{D},\mathbf{x}}\left[\left|h(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}\left[h(\mathbf{x};\mathcal{D}) \mid \mathbf{x}\right]\right|^2\right] + \mathbb{E}_{\mathcal{D},\mathbf{x}}\left[\left|\mathbb{E}_{\mathcal{D}}\left[h(\mathbf{x};\mathcal{D}) \mid \mathbf{x}\right] - y_*(\mathbf{x})\right|^2\right] + \\
&2\mathbb{E}_{\mathcal{D},\mathbf{x}}\left[\left(h(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}\left[h(\mathbf{x};\mathcal{D}) \mid \mathbf{x}\right]\right)\left(\mathbb{E}_{\mathcal{D}}\left[h(\mathbf{x};\mathcal{D}) \mid \mathbf{x}\right] - y_*(\mathbf{x})\right)\right] = \\
&\mathbb{E}_{\mathcal{D},\mathbf{x}}\left[\left|h(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}\left[h(\mathbf{x};\mathcal{D}) \mid \mathbf{x}\right]\right|^2\right] + \mathbb{E}_{\mathbf{x}}\left[\left|\mathbb{E}_{\mathcal{D}}\left[h(\mathbf{x};\mathcal{D}) \mid \mathbf{x}\right] - y_*(\mathbf{x})\right|^2\right]
\end{aligned}$$

- Try to convince yourself that the inner product term is zero.
- This is the bias and variance decomposition for the general estimator (with the squared error loss).

Bias-Variance Decomposition for the General Estimator

Bias-Variance Decomposition

$$\mathbb{E}_{\mathcal{D},\mathbf{x}}\left[\left|h(\mathbf{x};\mathcal{D}) - y_*(\mathbf{x})\right|^2\right] = \underbrace{\mathbb{E}_{\mathbf{x}}\left[\left|\mathbb{E}_{\mathcal{D}}\left[h(\mathbf{x};\mathcal{D}) \mid \mathbf{x}\right] - y_*(\mathbf{x})\right|^2\right]}_{\text{bias}} + \underbrace{\mathbb{E}_{\mathcal{D},\mathbf{x}}\left[\left|h(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}\left[h(\mathbf{x};\mathcal{D}) \mid \mathbf{x}\right]\right|^2\right]}_{\text{variance}}.$$

- Bias: The squared error between the average estimator (averaged over dataset $\mathcal{D}$) and the best predictor $y_*(\mathbf{x}) = \mathbb{E}\left[t|\mathbf{x}\right]$, averaged over $\mathbf{x} \sim p_{\mathbf{x}}$.
- Variance: The variance of a single estimator $h(\mathbf{x};\mathcal{D})$ (whose randomness comes from $\mathcal{D}$).
  - Note that $\mathbb{E}_{\mathcal{D},\mathbf{x}}\left[\left|h(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}\left[h(\mathbf{x};\mathcal{D}) \mid \mathbf{x}\right]\right|^2\right] = \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[\left|h(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}\left[h(\mathbf{x};\mathcal{D}) \mid \mathbf{x}\right]\right|^2\right]\right] = \mathbb{E}_{\mathbf{x}}\left[\mathrm{Var}_{\mathcal{D}}[h(\mathbf{x};\mathcal{D})|\mathbf{x}]\right].$

Bias-Variance Decomposition for the General Case

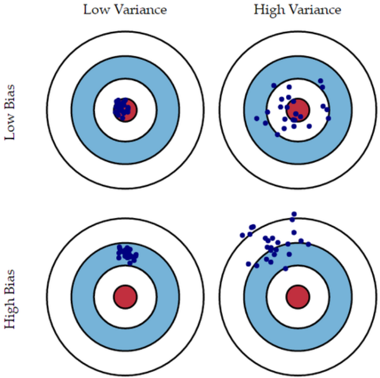
Bias-Variance Decomposition

$$\mathbb{E}_{\mathcal{D},\mathbf{x}}\left[|h(\mathbf{x};\mathcal{D}) - t|^2\right] = \underbrace{\mathbb{E}_{\mathbf{x}}\left[|\mathbb{E}_{\mathcal{D}}\left[h(\mathbf{x};\mathcal{D}) \mid \mathbf{x}\right] - y_*(\mathbf{x})|^2\right]}_{\text{bias}} + \underbrace{\mathbb{E}_{\mathcal{D},\mathbf{x}}\left[|h(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}\left[h(\mathbf{x};\mathcal{D}) \mid \mathbf{x}\right]|^2\right]}_{\text{variance}} + \underbrace{\mathbb{E}\left[|y_*(\mathbf{x}) - t|^2\right]}_{\text{Bayes error}}.$$

- We have an additional term of $\mathbb{E}\left[|y_*(\mathbf{x}) - t|^2\right] = \mathbb{E}_{\mathbf{x}}\left[\mathrm{Var}[t \mid \mathbf{x}]\right]$ (Why?!).
- This is due to the the variance of t at each fixed $\mathbf{x}$, averaged over $\mathbf{x} \sim p_{\mathbf{x}}$. As before, this comes from the randomness of the r.v. t and cannot be avoided. This is the Bayes error.

Bias-Variance Decomposition: A Visualization

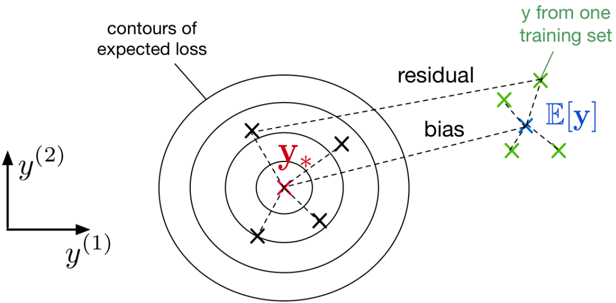
- Throwing darts = predictions for each draw of a dataset



- What doesn't this capture?
- We average over points **x** from the data distribution

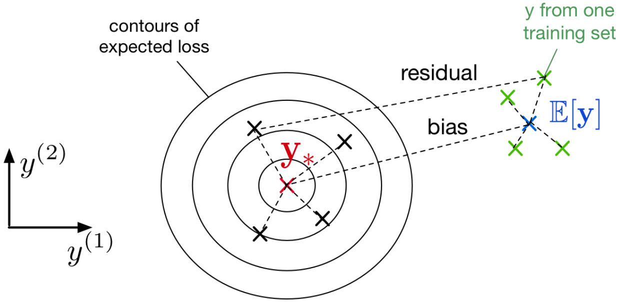
Bias-Variance Decomposition: Another Visualization

- We can visualize this decomposition in the output space, where the axes correspond to predictions on the test examples.
- Consider two test query inputs $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$. The outputs are
  - The Bayes optimal prediction $y_* = [y_*(\mathbf{x}^{(1)}), y_*(\mathbf{x}^{(2)})] = [\mathbb{E}\left[t|\mathbf{x}^{(1)}\right], \mathbb{E}\left[t|\mathbf{x}^{(2)}\right]]$.
  - The prediction of the model $h(\mathbf{x}; \mathcal{D})$: $y = [y^{(1)}, y^{(2)}] = [h(\mathbf{x}^{(1)}; \mathcal{D}), h(\mathbf{x}^{(2)}; \mathcal{D})]$.
- We can visualize the outputs as follows:



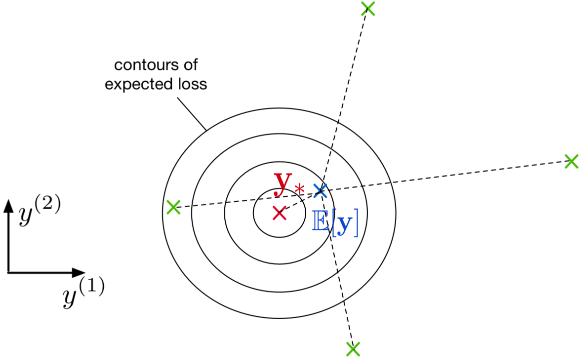
Bias-Variance Decomposition: Another Visualization

- If we have an overly simple model (e.g., K-NN with large K), it might have
  - high bias (because it is too simplistic to capture the structure in the data)
  - low variance (because there is enough data to get a stable estimate of the decision boundary)



Bias-Variance Decomposition: Another Visualization

- If you have an overly complex model (e.g., K-NN with $K = 1$), it might have
  - low bias (since it learns all the relevant structure)
  - high variance (it fits the quirks of the data you happened to sample)



Ensemble Methods – Part I: Bagging

Ensemble Methods: Brief Overview

- An ensemble of predictors is a set of predictors whose individual decisions are combined in some way to predict new examples, for example by (weighted) majority vote.
- For the result to be nontrivial, the learned hypotheses must differ somehow, for example because of
  - Trained on different data sets
  - Trained with different weighting of the training examples
  - Different algorithms
  - Different choices of hyperparameters
- Ensembles are usually easy to implement. The hard part is deciding what kind of ensemble you want, based on your goals.
- Two major types of ensembles methods:
  - Bagging
  - Boosting

Bagging: Motivation

- Suppose that we could somehow sample m independent training sets $\{\mathcal{D}_i\}_{i=1}^m$ from p_{dataset}.
- We could then learn a predictor $h_i \triangleq h(\cdot; \mathcal{D}_i)$ based on each dataset, and take the average $h(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^m h_i(\mathbf{x})$.
- How does this affect the terms of the expected loss?
  - **Bias: Unchanged**, since the averaged prediction has the same expectation

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_i,\ldots,\mathcal{D}_m \overset{\text{i.i.d.}}{\sim} p_{\text{dataset}}}\left[h(\mathbf{x})\right] &= \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{\mathcal{D}_i \sim p_{\text{dataset}}}\left[h_i(\mathbf{x})\right] \\ &= \mathbb{E}_{\mathcal{D} \sim p_{\text{dataset}}}\left[h(\mathbf{x}; \mathcal{D})\right].\end{aligned}$$

  - **Variance: Reduced**, since we are averaging over independent samples

$$\underset{\mathcal{D}_1,\ldots,\mathcal{D}_m}{\mathrm{Var}}[h(\mathbf{x})] = \frac{1}{m^2}\sum_{i=1}^m \underset{\mathcal{D}_i}{\mathrm{Var}}[h_i(\mathbf{x})] = \frac{1}{m}\underset{\mathcal{D}}{\mathrm{Var}}[h_{\mathcal{D}}(\mathbf{x})].$$

- Q: What if $m \to \infty$?

Bagging

- In practice, we do not have access to the underlying data generating distribution p_{sample}.
- It is expensive to collect many i.i.d. datasets from p_{dataset}.
- Solution: bootstrap aggregation, or bagging.
    - Take a single dataset $\mathcal{D}$ with n examples.
    - Generate m new datasets, each by sampling n training examples from $\mathcal{D}$, with replacement.
    - Average the predictions of models trained on each of these datasets.
- Bagging works well for low-bias / high-variance estimators.

Bagging

- Problem: the datasets are not independent, so we do not get the $\frac{1}{m}$ variance reduction.
- Possible to show that if the sampled predictions have variance σ^2 and correlation ρ, then

$$\mathrm{Var}\left(\frac{1}{m}\sum_{i=1}^{m} h_i(\mathbf{x})\right) = \rho\sigma^2 + \frac{1}{m}(1-\rho)\sigma^2.$$

  - Exercise: Prove this! (See next slide)
- By increasing m, the second term decreases.
- The first term, however, remains the same. It limits the benefit of bagging.
- If we can make correlation ρ as small as possible, we benefit more from bagging.

Some Properties of Variance

- Covariance:

$$\mathbf{Cov}\left(X,Y\right)=\mathbb{E}\left[(X-\mathbb{E}\left[X\right])(Y-\mathbb{E}\left[Y\right])\right].$$

- Correlation:

$$\rho_{X,Y}=\frac{\mathbf{Cov}\left(X,Y\right)}{\sigma_X\sigma_Y}.$$

- Covariance of linear combination:

$$\begin{aligned}\text{Var}\left[\sum_{i=1}^{m}Z_i\right]&=\sum_{i,j=1}^{m}\mathbf{Cov}\left(Z_i,Z_j\right)\\&=\sum_{i=1}^{m}\text{Var}[Z_i]+\sum_{i,j=1;i\neq j}^{m}\mathbf{Cov}\left(Z_i,Z_j\right).\end{aligned}$$

Bagging

$$\mathrm{Var}\left(\frac{1}{m}\sum_{i=1}^{m} h_i(\mathbf{x})\right) = \rho\sigma^2 + \frac{1}{m}(1-\rho)\sigma^2.$$

- It can be advantageous to introduce *additional* variability into your algorithm, as long as it reduces the correlation between samples.
  - Intuition: you want to invest in a diversified portfolio, not just one stock.
  - Can help to use average over multiple algorithms, or multiple configurations (i.e., hyperparameters) of the same algorithm.

Random Forests

- Random forests: bagged decision trees, with one extra trick to decorrelate the predictions
- When choosing each node of the decision tree, choose a random set of p input attributes (e.g., $p = \sqrt{d}$), and only consider splits on those features.
  - Smaller p reduces the correlation between trees.
- Random forests improve the variance reduction of bagging by reducing the correlation between the trees (ρ).
- For regression, we take the average output of the ensemble; for classification, we perform a majority vote.
- Random forests are probably one of the best black-box machine learning algorithm. They often work well with no tuning whatsoever.
  - One of the most widely used algorithms in Kaggle competitions.

Conclusion

- Bias-Variance Decomposition
  - The error of a machine learning algorithm can be decomposed to a bias term and a variance term.
  - Hyperparameters of an algorithm might allow us to tradeoff between these two.
- Ensemble Methods
  - Bagging as a simple way to reduce the variance of an estimation method