Bagging
Ensemble Methods - Classification

- **Problem**: given $T$ binary classification hypotheses $(h_1, \ldots, h_T)$ find combined classifier

$$f : x \mapsto \text{sgn} \left( \sum_{t=1}^{T} \alpha_t h_t \right),$$

with better performance.

- When does it work? Need diversity (e.g., different features, different training sets).

  - use different subsets of the data for training.
Bagging - Classification

Bagging (Bootstrap aggregating).

\[ \text{Bagging}(S = ((x_1, y_1), \ldots, (x_m, y_m))) \]

1. \textbf{for} \( t \leftarrow 1 \) \textbf{to} \( T \) \textbf{do}
2. \( S_t \leftarrow \text{Bootstrap}(S) \) \( \triangleright \) i.i.d. sampling with replacement from \( S \).
3. \( h_t \leftarrow \text{TrainClassifier}(S_t) \)
4. \textbf{return} \( h_S = x \mapsto \text{MajorityVote}((h_1(x), \ldots, h_T(x))) \)
Bagging - Regression

Bagging (Bootstrap aggregating).

\hspace{.5em} \text{Bagging} (\text{S} = ((x_1, y_1), \ldots, (x_m, y_m)))

1 \hspace{.5em} \textbf{for} \ t \leftarrow 1 \ \textbf{to} \ T \ \textbf{do}
2 \hspace{.5em} S_t \leftarrow \text{BOOTSTRAP}(S) \triangleright \text{i.i.d. sampling with replacement from } S.
3 \hspace{.5em} h_t \leftarrow \text{TRAIN\textsc{RegressionAlgorithm}}(S_t)
4 \hspace{.5em} \textbf{return} \ h_S = x \mapsto \text{MEAN}((h_1(x), \ldots, h_T(x)))
Bias-Variance Decomposition

\textbf{Proposition:} for any hypothesis $h_S$, the following decomposition holds:

\[
\mathbb{E}_S \left[ \mathbb{E}_{X,Y} [(h_S(X) - Y)^2] \right] \\
= \mathbb{E}_{S,X} \left[ (h_S(X) - \mathbb{E}_S[h_S(X)])^2 \right] + \mathbb{E}_X \left[ (\mathbb{E}_S[h_S(X)] - \mathbb{E}[Y|X])^2 \right] + \mathbb{E}_X \left[ (Y - \mathbb{E}[Y|X])^2 \right].
\]

\textbf{Bias-variance minimization trade-off:}
\begin{itemize}
  \item small $S$ and large $H$: small bias, large variance.
  \item large $S$ and small $H$: large bias, small variance.
\end{itemize}
Bias-Variance Decomp. Proof

- Observe that

\[
E_{X,Y} [(h_S(X) - E[Y|X])(E[Y|X] - Y)]
= E_X \left[ E_{Y|X} [(h_S(X) - E[Y|X])(E[Y|X] - Y)] \right]
= E_X \left[ (h_S(X) - E[Y|X]) E_{Y|X} [(E[Y|X] - Y)] \right] = 0.

- Therefore,

\[
E_{X,Y} [(h_S(X) - Y)^2] = E_{X,Y} \left[ [(h_S(X) - E[Y|X]) + (E[Y|X] - Y)]^2 \right]
= E_X [(h_S(X) - E[Y|X])^2] + E_{X,Y} [(E[Y|X] - Y)^2].
\]
Bias-Variance Decomp. Proof

Since \( \mathbb{E}_S \left[ h_S(X) - \mathbb{E}_S[h_S(X)] \right] = 0 \), the following holds:

\[
\mathbb{E}_S \left[ \mathbb{E}_{X,Y} \left[ (h_S(X) - \mathbb{E}[Y|X])^2 \right] \right] = \mathbb{E}_X \left[ \mathbb{E}_S \left[ (h_S(X) - \mathbb{E}_S[h_S(X)])^2 \right] \right] + \mathbb{E} \left[ \mathbb{E}_S \left[ (\mathbb{E}_S[h_S(X)] - \mathbb{E}[Y|X])^2 \right] \right].
\]
Ensemble Bias

- Bias of averaged hypothesis in regression:

\[
\text{Bias} (h_S, x) = \mathbb{E}_{S \sim D^m} \left[ \frac{1}{T} \sum_{t=1}^{T} h_t(x) \right] - \mathbb{E}[Y|x]
\]

\[
= \frac{1}{T} \sum_{t=1}^{T} \left[ \mathbb{E}_{S \sim D^m} [h_t(x)] - \mathbb{E}[Y|x] \right]
\]

\[
= \frac{1}{T} \sum_{t=1}^{T} \text{Bias} (h_t, x).
\]

Thus, relatively unbiased base hypotheses lead to relatively unbiased ensemble.
Ensemble Variance

**Proposition:** for any \( x \), the variance of the ensemble hypothesis at \( x \) is given by

\[
\text{Var} (h_S, x) = \frac{1}{T^2} \sum_{t=1}^{T} \text{Var} (h_t, x) + \frac{1}{T^2} \sum_{t \neq t'} \text{Cov}[h_t(x), h_{t'}(x)].
\]

- thus, if approximately uncorrelated base hypotheses \( \text{Cov}[h_t(x), h_{t'}(x)] \approx 0 \).
- assume approximately equal variances, then

\[
\text{Var} (h_S, x) \approx \frac{1}{T} \text{Var} (h_1, x).
\]

reduction by \( 1/T \).
Bagging - Regression

Regression properties:
- small covariances (different subsets).
- similar variances (on average).
- similar biases.

Classification: unclear explanation.
References

