Statistical Learning, Fall 2024-5

Class notes 10

The gradient boosting paradigm

- Choose a loss function for modeling (like RSS for regression)
- At each iteration: calculate the (negative) gradient of the loss function at the current model, use that as $Y^{(t)}$ for the next weak learner
- Interpretation: trying to find a weak learner h_{k_t} which "behaves like" the negative gradient, which is the direction of fastest decrease of the loss
- Can be applied with different loss functions for regression or classification

For 2-class classification we denote $y_i \in \{0, 1\}$ (sometimes $y_i \in \{\pm 1\}$ but we stick here with 0/1). A common loss function, which we also presented in the context of logistic regression is the (negative) Bernoulli log likelihood:

$$L(y_i, \hat{y}_i) = -y_i \log(\frac{\exp(\hat{y}_i)}{1 + \exp(\hat{y}_i)}) - (1 - y_i) \log(\frac{1}{1 + \exp(\hat{y}_i)}).$$

For simplicity denote as in logistic regression using the inverse logit transformation:

$$\hat{p}_i = \frac{\exp(\hat{y}_i)}{1 + \exp(\hat{y}_i)} = \frac{\exp(F^{(t-1)}(x_i))}{1 + \exp(F^{(t-1)}(x_i))},$$

where the last equality refers to already applying $\hat{y}_i = F^{(t-1)}(x_i)$.

If we use this loss function in a gradient boosting algorithm, after some scary differentiation we get that simply:

$$y_i^{(t)} = y_i(1 - \hat{p}_i) - (1 - y_i)\hat{p}_i.$$

Note that since $y_i \in \{0, 1\}$, only one of the two expressions is non-zero.

Formal gradient boosting description

In the gradient boosting paradigm, define:

- Training loss function per observation: $L(y, \hat{y})$, and total: $\mathcal{L} = \sum_i L(Y_i, \hat{Y}_i)$.
- Family of q weak learners: $h_k : \mathcal{X} \to \mathbb{R}$, $k = 1, \ldots, q$, possibly $q = \infty$.

Now define $F^{(0)} \equiv 0$, and for $t = 1, \ldots, T$:

1. Set $\nabla \mathcal{L} = \left\{ \frac{\partial L(y_i, \hat{y}_i)}{\partial \hat{y}_i}\Big|_{\hat{y}_i = F^{(t-1)}(x_i)} \right\}_{i=1}^{n}$
2. Solve (exactly or approximately) $k_t = \arg\min_k \langle \nabla \mathcal{L}, h_k(\mathbb{X}) \rangle$
3. Find coefficient $\alpha_t = \begin{cases} \arg\min_\alpha \mathcal{L}(F^{(t-1)} + \alpha h_k) & \text{Line search boosting} \\ \epsilon \text{ (small)} & \epsilon\text{-boosting} \end{cases}$
4. Update $F^{(t)} = F^{(t-1)} + \alpha_t h_{k_t}$.

The optimization problem in the second step is often replaced with:

$$k_t = \arg\min_k \|(-\nabla \mathcal{L}) - h_k(\mathbb{X})\|^2$$

which is very similar, except it also penalizes $\|h_k(\mathbb{X})\|^2$, controlling the norm of the model.

Gradient boosting and AdaBoost

The famous AdaBoost algorithm (Freund and Schapire 1992) was developed in the machine learning community, with a different theory, but we can discuss it as gradient boosting. The algorithm for two-class classification initializes $w_i \equiv 1$, then for $t = 1...T$ updates:

1. Fit a classification tree with response y and weights w on the observations, getting tree h_t
2. Denote by Err_t the (weighted) misclassification error of h_t
3. Set $\alpha_t = 0.5 \log((1 - Err_t)/Err_t)$
4. Update weights: $w_i \leftarrow w_i \exp(-\alpha_t (y_i h_t(x_i)))$

This is a gradient boosting algorithm with $L(y, \hat{y}) = \exp(-y\hat{y})$, (where $y \in \{\pm 1\}$ and $\hat{y} \in \mathbb{R}$). The weights w are the gradient, and step 3 is the solution to line search in this setting.

Model evaluation and selection

The possible goals:

- Model evaluation: estimate prediction error for a given model (or modeling approach "black box")
- Model selection: among a set of candidate modeling approaches, seek to select one with low prediction error

Definitions of prediction error, given a prediction loss function $L(y, \hat{y})$:

- Fixed-X error: Assume $\mathbb{X}$ given and fixed and shared between training and prediction, define the prediction error: $\frac{1}{n}\mathbb{E}_{\mathbb{Y}, \mathbb{Y}^{new}} \sum_i L(Y_i^{new}, \hat{Y}_i)$.
- Same-X error: Similar to Fixed-X, but assume $\mathbb{X}$ is random: $\mathbb{E}_{\mathbb{X}} \left[\frac{1}{n}\mathbb{E}_{\mathbb{Y}, \mathbb{Y}^{new}} \sum_i L(Y_i^{new}, \hat{Y}_i) | \mathbb{X} \right]$.
- Random-X error: The realistic scenario when new X is used for prediction:

$$\mathbb{E}_{T=(\mathbb{X},\mathbb{Y}),X,Y} \sum_i L(Y, \hat{f}(X)).$$

Validation and cross validation

For a given data set T of size n, we can divide it into two or three pieces for training, validation (model selection) and test (model evaluation). A typical split is $60\% - 20\% - 20\%$. If we want only validation or only test we can split $80\% - 20\%$.

A more advanced approach uses K-fold cross-validation (CV): define a random partition of the n data points into K sets, each of size n/K (exactly or approximately). Express it as a function:

$$\phi : \{1, \ldots, n\} \to \{1, \ldots, K\},$$

and call the set $\phi_k = \{i : \phi(i) = k\}$ the kth fold.

Model evaluation using cross validation: for $k = 1, \ldots, K$:

- Build a model $\hat{f}^{(k)}$ using all folds except the kth fold (total $(K-1)/K \cdot n$ data points)
- $L_k = \sum_{i \in \phi_k} L(Y_i, \hat{f}^{(k)}(X_i))$

And define the CV estimate: $CV = \frac{1}{n} \sum_{k=1}^{K} L_k$. This way, the model is evaluated on the entire n observations. Each model is built on $(K-1)/K \cdot n < n$ observations, and we generally want to make K as large as possible, to get a realistic evaluation of the performance of the model built on n observations. However making K too large can have substantial computational cost.

Model selection using CV: Assume the modeling approach has a tuning parameter λ, then we would calculate:

$$L_k(\lambda) = \sum_{i \in \phi_k} L(Y_i, \hat{f}_\lambda^{(k)}(X_i)) \ , \ \ CV(\lambda) = \frac{1}{n} \sum_{k=1}^{K} L_k(\lambda) \ , \ \ \hat{\lambda} = \arg\min_\lambda CV(\lambda),$$

then we rebuild the model on the entire n observations using $\hat{\lambda}$.

n-fold (Leave-one-out) CV and the leaving out lemma

The Leaving Out Lemma requires two conditions under the squared loss, iid error assumption:

1. Linear model: $\hat{\mathbb{Y}} = S(\mathbb{X})\mathbb{Y}$ in training.
2. For any $1 \le i_0 \le n$, define a pseudo training dataset $\mathbb{X}, \tilde{\mathbb{Y}}$ with the same $\mathbb{X}$ as our training data, and $\tilde{y}_j = y_j$ for $j \neq i_0$ and:
$$\tilde{y}_{i_0} = \hat{y}_{i_0}^{(-i_0)},$$
where the superscript $(-i_0)$ indicates the model built on $n-1$ observations, leaving out i_0. Then we require:
$$\hat{\tilde{y}}_{i_0} = (S\tilde{\mathbf{y}})_{i_0} = \hat{y}_{i_0}^{(-i_0)}.$$

Under these conditions we can easily prove (on the board) that:

$$(y_{i_0} - \hat{y}_{i_0}^{(-i_0)}) = \frac{(y_{i_0} - \hat{y}_{i_0})}{1 - S_{i_0 i_0}},$$

and therefore:

$$LOOCV = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{(1 - S_{ii})^2},$$

and we can calculate LOOCV by only fitting the model on the training data once.

We show in class the least squares regression complies with both conditions, and stated that ridge regression also does. K-NN complies with the first but not the second, while we stated that Lasso does not comply with the result, but left it as a challenge to figure out which of the conditions do not hold.

Optimism and degrees of freedom

Recall for Fixed-X, we have the training squared loss: $\frac{1}{n}RSS = \frac{1}{n}\|\mathbb{Y} - \hat{\mathbb{Y}}\|^2$,
and the prediction lost for $\mathbb{Y}^{new}$ an iid copy of $\mathbb{Y}$ at same $\mathbb{X}$: "EPE" $= \mathbb{E}_{\mathbb{Y},\mathbb{Y}^{new}} \frac{1}{n}\|\mathbb{Y}^{new} - \hat{\mathbb{Y}}\|^2$ (where $\mathbb{Y}$ plays a role through defining $\hat{\mathbb{Y}}$).

It is natural to define the *optimism* of a model building approach (a mapping $\mathbb{Y} \to \hat{\mathbb{Y}}$) as the difference in expectation between the two measures:

$$op = \mathbb{E}_{\mathbb{Y},\mathbb{Y}^{new}} \frac{1}{n}\|\mathbb{Y}^{new} - \hat{\mathbb{Y}}\|^2 - \frac{1}{n}\|\mathbb{Y} - \hat{\mathbb{Y}}\|^2.$$

The beautiful and fundamental result is that under very general assumptions we have:

$$op = \frac{2}{n}\sum_{i=1}^{n} Cov(y_i, \hat{y}_i).$$

Furthermore, we can actually calculate or estimate this quantity for many models of interest. If we can calculate op or an estimate $\hat{op}$, then we can use it to obtain an unbiased estimate of the prediction error as :

$$\frac{1}{n}RSS + op.$$

Proof of optimism formula:

$$\begin{aligned}
\mathbb{E}\|\mathbb{Y} - \hat{\mathbb{Y}}\|^2 &= \mathbb{E}\|\mathbb{Y} - \mathbb{E}\mathbb{Y} + \mathbb{E}\mathbb{Y} - \mathbb{E}\hat{\mathbb{Y}} + \mathbb{E}\hat{\mathbb{Y}} - \hat{\mathbb{Y}}\|^2 = \mathbb{E}\|\mathbb{Y} - \mathbb{E}\mathbb{Y}\|^2 + \|\mathbb{E}\mathbb{Y} - \mathbb{E}\hat{\mathbb{Y}}\|^2 + \mathbb{E}\|\mathbb{E}\hat{\mathbb{Y}} - \hat{\mathbb{Y}}\|^2 + \\
&+ \underbrace{2\mathbb{E}(\mathbb{Y} - \mathbb{E}\mathbb{Y})^T(\mathbb{E}\mathbb{Y} - \mathbb{E}\hat{\mathbb{Y}})}_{A} + \underbrace{2\mathbb{E}(\mathbb{Y} - \mathbb{E}\mathbb{Y})^T(E\hat{\mathbb{Y}} - \hat{\mathbb{Y}})}_{B} + \underbrace{2(\mathbb{E}\mathbb{Y} - \mathbb{E}\hat{\mathbb{Y}})^T\mathbb{E}(\mathbb{E}\hat{\mathbb{Y}} - \hat{\mathbb{Y}})}_{C} = \\
&= \underbrace{\|\mathbb{Y} - \mathbb{E}\mathbb{Y}\|^2}_{\text{Irreducible}} + \underbrace{\|\mathbb{E}\mathbb{Y} - \mathbb{E}\hat{\mathbb{Y}}\|^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}\|\mathbb{E}\hat{\mathbb{Y}} - \hat{\mathbb{Y}}\|^2}_{\text{Variance}} - 2\underbrace{\mathbb{E}(\mathbb{Y} - \mathbb{E}\mathbb{Y})^T(\hat{\mathbb{Y}} - \mathbb{E}\hat{\mathbb{Y}})}_{\text{Covariance}}
\end{aligned}$$

It is easy to see that $A = C = 0$ from arguments we have seen previously. For the prediction error (remember that $\mathbb{Y}^{new}, \mathbb{Y}$ are identically distributed and in particular $\mathbb{E}\mathbb{Y}^{new} = \mathbb{E}\mathbb{Y}$):

$$\begin{aligned}
\mathbb{E}\|\mathbb{Y}^{new} - \hat{\mathbb{Y}}\|^2 &= \mathbb{E}\|\mathbb{Y}^{new} - \mathbb{E}\mathbb{Y} + \mathbb{E}\mathbb{Y} - \mathbb{E}\hat{\mathbb{Y}} + \mathbb{E}\hat{\mathbb{Y}} - \hat{\mathbb{Y}}\|^2 = \mathbb{E}\|\mathbb{Y} - \mathbb{E}\mathbb{Y}\|^2 + \|\mathbb{E}\mathbb{Y} - \mathbb{E}\hat{\mathbb{Y}}\|^2 + \mathbb{E}\|\mathbb{E}\hat{\mathbb{Y}} - \hat{\mathbb{Y}}\|^2 + \\
&+ \underbrace{2\mathbb{E}(\mathbb{Y}^{new} - \mathbb{E}\mathbb{Y})^T(\mathbb{E}\mathbb{Y} - \mathbb{E}\hat{\mathbb{Y}})}_{A} + \underbrace{2\mathbb{E}(\mathbb{Y}^{new} - \mathbb{E}\mathbb{Y})^T(E\hat{\mathbb{Y}} - \hat{\mathbb{Y}})}_{B} + \underbrace{2(\mathbb{E}\mathbb{Y} - \mathbb{E}\hat{\mathbb{Y}})^T\mathbb{E}(\mathbb{E}\hat{\mathbb{Y}} - \hat{\mathbb{Y}})}_{C} = \\
&= \underbrace{\|\mathbb{Y} - \mathbb{E}\mathbb{Y}\|^2}_{\text{Irreducible}} + \underbrace{\|\mathbb{E}\mathbb{Y} - \mathbb{E}\hat{\mathbb{Y}}\|^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}\|\mathbb{E}\hat{\mathbb{Y}} - \hat{\mathbb{Y}}\|^2}_{\text{Variance}}
\end{aligned}$$

And it follows that:

$$op = \frac{1}{n}\left(\mathbb{E}\|\mathbb{Y}^{new} - \hat{\mathbb{Y}}\|^2 - \mathbb{E}\|\mathbb{Y} - \hat{\mathbb{Y}}\|^2\right) = \frac{2}{n}\sum_{i=1}^{n} Cov(y_i, \hat{y}_i) = \frac{2}{n} tr(Cov(\mathbb{Y}, \hat{\mathbb{Y}})).$$

The simple setting where we can use this result is when:

- We have iid error model $y = f(X) + \epsilon\ ,\ \epsilon \sim (0, \sigma^2)$
- We have a linear model (in the generalized sense): $\hat{\mathbb{Y}} = S(\mathbb{X})\mathbb{Y}$.

(notice no linearity or normality assumptions). In these cases we can write:

$$op = \frac{2}{n} tr(Cov(\mathbb{Y}, S\mathbb{Y})) = \frac{2}{n} tr(SCov(\mathbb{Y}, \mathbb{Y})) = \frac{2\sigma^2}{n} tr(S\mathbb{I}) = \frac{2\sigma^2 tr(S)}{n}.$$

For least squares: $S = H = \mathbb{X}\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T$, so:

$$op = \frac{2\sigma^2}{n} tr(H) = \frac{2p}{n}\sigma^2,$$

in other words, an unbiased estimate of Fixed-X prediction error is:

$$\frac{RSS(\hat{\beta})}{n} + \frac{2p\sigma^2}{n}.$$

For ridge regression: $S_\lambda = \mathbb{X}\left(\mathbb{X}^T\mathbb{X} + \lambda\mathbb{I}\right)^{-1}\mathbb{X}^T$ for $\lambda > 0$, so:

$$op = \frac{2\sigma^2}{n} tr(S_\lambda) < \frac{2p}{n}\sigma^2,$$

and we see the reduced optimism from adding regularization.

We can also apply this result to k-NN which has the required form for Fixed-X, the result? in HW4...

Interesting extensions we may discuss as time permits:

- Dealing with unknown σ^2 (e.g. using unbiased estimates assuming linear model)
- Extending beyond squared loss to likelihood-loss (as in logistic regression): AIC
- Extending to cases where *op* cannot be calculated but can be estimated in unbiased manner: Stein's Lemma