HW1

Brian Allen,ba2542
Serena Zhang,mz2642
Haozheng Ni, hn2318

February 13, 2018

1.

2. (a) The test error should be the same. At every split, greedy decision tree will search all threshold for each variable. This will not be affected by monotonic transformation of features. For example, at certain split, C_0 splits some feature X_i with threshold t_i. Assume the original mean of feature X_i of all the sample is μ_i, then the new threshold $t_i - \mu_i$ on centered feature $\hat{X}_i$ will give the same classification result, i.e the gini index is the same with C_0.

   (b) The test error should be the same. Similarly as above question, the new threshold becomes $\frac{t_i - \mu_i}{\sigma_i}$ where σ_i^2 is the variance of feature X_i, and this gives the same classification result as C_0.

   (c) The test error should be the same. Consider the distance of a new data $\mathbf{x}$ with a training data $\mathbf{x}_i$. Denote $\mathbf{x} = [x_1 x_2 ... x_k]^T, \mathbf{x}_i = [x_{i1} x_{i2} ... x_{ik}]^T$.

$$d(\mathbf{x}, \mathbf{x}_i) = (x_1 - x_{i1})^2 + ... + (x_k - x_{ik})^2$$

   Assume now we centralize all the features, the means of each feature in training data are $\mu_1, .., \mu_k$, and the centered $\mathbf{x}, \mathbf{x}_i$ are denoted by $\hat{\mathbf{x}}, \hat{\mathbf{x}}_i$. Then we have

$$\begin{aligned} d(\hat{\mathbf{x}}, \hat{\mathbf{x}}_i) &= ((x_1 - \mu_1) - (x_{i1} - \mu_1))^2 + ... + ((x_k - \mu_k) - (x_{ik} - \mu_k))^2 \\ &= d(\mathbf{x}, \mathbf{x}_i) \end{aligned}$$

   Therefore the pairwise distance is not changed by centralization, which means the classification result remains the same.

   (d) The test errors should be different. Denote sample variance of each feature as $\sigma_1^2, .., \sigma_k^2$, then the new distance between standardized $\hat{\mathbf{x}}, \hat{\mathbf{x}}_i$ is

$$\begin{aligned} d(\hat{\mathbf{x}}, \hat{\mathbf{x}}_i) &= (\frac{(x_1 - \mu_1)}{\sigma_1} - \frac{(x_{i1} - \mu_1)}{\sigma_1})^2 + ... + (\frac{(x_k - \mu_k)}{\sigma_k} - \frac{(x_{ik} - \mu_k)}{\sigma_k})^2 \\ &= \frac{1}{\sigma_1^2}((x_1 - \mu_1) - (x_{i1} - \mu_1))^2 + ... + \frac{1}{\sigma_k^2}((x_k - \mu_k) - (x_{ik} - \mu_k))^2 \end{aligned}$$

   If $\sigma_i's$ are not all 1, above distance does not necessarily equal to $d(\mathbf{x}, \mathbf{x}_i)$. Therefore the nearest neighbor for a new sample $\mathbf{x}$ could be different. Thus the test error should be different.

3. First we want to show empirical risk is an unbiased estimator of risk, i.e For a given iid sample $(X_1, Y_1), ..., (X_n, Y_n)$, and an arbitrary estimator β we have

$$R(\beta) = E[\frac{1}{n}\sum_{i=1}^{n}(X_i^T\beta - Y_i)^2]$$

Proof.

$$\begin{aligned}
E[\frac{1}{n}\sum_{i=1}^{n}(X_i^T\beta - Y_i)^2] &= \frac{1}{n}\sum_{i=1}^{n}E[(X_i^T\beta - Y_i)^2] \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_{X,Y}(X_i^T\beta - Y_i)^2 p(X_i, Y_i)dX_i dY_i \\
&= \frac{1}{n}\sum_{i=1}^{n}R(\beta) \qquad \text{(because of iid property)} \\
&= R(\beta)
\end{aligned}$$

□

To make the notation simple, the ERM is written as

$$\hat{R}(\beta) = \frac{1}{n}\|\mathbf{X}\beta - \mathbf{Y}\|^2$$

where $\mathbf{X}$ is the design matrix of an iid sample D. Moreover, since the ERM estimator of β depends on the sample, we can write the estimator $\hat{\beta}$ give an arbitrary sample $D = (\mathbf{X}, \mathbf{Y})$ by $\hat{\beta} = \beta(\mathbf{X}, \mathbf{Y})$. Then we have $\tilde{\beta} = \beta(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})$.
Then we'd like to show given two iid sample $D, \tilde{D}$, and we find ERM estimator of β for each sample, denoted by $\hat{\beta}, \tilde{\beta}$ the expected empirical risk is the same, i.e

$$E(\hat{R}(\hat{\beta})) = E(\tilde{R}(\tilde{\beta}))$$

Notice that the expectations are taken on sample $D, \tilde{D}$ respectively for each side.

Proof.

$$E(\hat{R}(\hat{\beta})) = \frac{1}{n}\int_D \|\mathbf{X}\beta(\mathbf{X}, \mathbf{Y}) - \mathbf{Y}\|^2 P(\mathbf{X}, \mathbf{Y})\, d\mathbf{X}d\mathbf{Y}$$

Since $D, \tilde{D}$ are iid sample, $(\mathbf{X}, \mathbf{Y})$ and $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})$ follow the same distribution. Therefore above equation can be written as

$$\begin{aligned}
E(\hat{R}(\hat{\beta})) &= \frac{1}{n}\int_{\tilde{D}} \|\tilde{\mathbf{X}}\beta(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) - \tilde{\mathbf{Y}}\|^2 P(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})\, d\tilde{\mathbf{X}}d\tilde{\mathbf{Y}} \\
&= \frac{1}{n}\int_{\tilde{D}} \|\tilde{\mathbf{X}}\tilde{\beta} - \tilde{\mathbf{Y}}\|^2 P(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})\, d\tilde{\mathbf{X}}d\tilde{\mathbf{Y}} \\
&= E(\tilde{R}(\tilde{\beta}))
\end{aligned}$$

□

Then we are ready to show

$$E(\tilde{R}(\hat{\beta})) \geq E(\hat{R}(\hat{\beta}))$$

Notice that we first fit $\hat{\beta}$ on sample D, and therefore the expectation on the left hand side is taken on both D ,and $\tilde{D}$, while that on the right is takes on D only.

Proof.

$$
\begin{aligned}
E(\tilde{R}(\hat{\beta})) &= \frac{1}{n}\int_D\int_{\tilde{D}} \|\tilde{X}\beta(\mathbf{X},\mathbf{Y}) - \tilde{\mathbf{Y}}\|^2 P(\tilde{\mathbf{X}},\tilde{\mathbf{Y}})P(\mathbf{X},\mathbf{Y})\, d\tilde{\mathbf{X}}d\tilde{\mathbf{Y}}d\mathbf{X}d\mathbf{Y} \\
&\geq \frac{1}{n}\int_D\int_{\tilde{D}} \|\tilde{X}\beta(\tilde{\mathbf{X}},\tilde{\mathbf{Y}}) - \tilde{\mathbf{Y}}\|^2 P(\tilde{\mathbf{X}},\tilde{\mathbf{Y}})P(\mathbf{X},\mathbf{Y})\, d\tilde{\mathbf{X}}d\tilde{\mathbf{Y}}d\mathbf{X}d\mathbf{Y} \\
&\quad \text{Because } \beta(\tilde{\mathbf{X}},\tilde{\mathbf{Y}}) \text{ minimizes the empirical risk on } \tilde{D} \\
&= \frac{1}{n}\int_D\int_{\tilde{D}} \tilde{R}(\tilde{\beta})P(\tilde{\mathbf{X}},\tilde{\mathbf{Y}})\, d\tilde{\mathbf{X}}d\tilde{\mathbf{Y}}P(\mathbf{X},\mathbf{Y})d\mathbf{X}d\mathbf{Y} \\
&= \frac{1}{n}\int_D E(\tilde{R}(\tilde{\beta}))P(\mathbf{X},\mathbf{Y})d\mathbf{X}d\mathbf{Y} \\
&= E(\tilde{R}(\tilde{\beta})) \qquad \text{As the integrand is independent of } D \\
&= E(\hat{R}(\hat{\beta})) \qquad \text{As proved above}
\end{aligned}
$$

□

As we have shown at very beginning, the empirical risk is an unbiased estimator of risk, we have

$$E(R(\hat{\beta})) = E(\tilde{R}(\hat{\beta})) \geq E(\hat{R}(\hat{\beta}))$$

4. (a)

$$
\begin{aligned}
E(\mathbf{Y}) &= \sum_{i=1}^{k} \mathbf{e}_i P(\mathbf{Y} = \mathbf{e}_i) \\
&= \sum_{i=1}^{k} \mathbf{e}_i p_i \\
&= [p_1\ p_2\ ...\ p_k]^T
\end{aligned}
$$

(b)

$$
\begin{aligned}
Cov(\mathbf{Y}) &= E[(\mathbf{Y} - E(\mathbf{Y}))(\mathbf{Y} - E(\mathbf{Y}))^T] \\
&= E(\mathbf{Y}\mathbf{Y}^T - E(\mathbf{Y})\mathbf{Y}^T - \mathbf{Y}E(\mathbf{Y})^T + E(\mathbf{Y})E(\mathbf{Y})^T)
\end{aligned}
$$

Let $E(\mathbf{Y}) = [p_1\ p_2\ ...\ p_k]^T, \mathbf{Y} = [y_1\ y_2\ ...\ y_k]^T$, the diagonal entry of above covariance matrix is

$$Cov(\mathbf{Y})_{ii} = E(y_i^2 - 2p_i y_i + p_i^2)$$

Since $P(y_i = 1) = p_i, P(y_i = 0) = 1 - p_i$, we have

$$E(y_i) = E(y_i^2) = p_i$$

Therefore

$$Cov(\mathbf{Y})_{ii} = p_i - 2p_i^2 + p_i^2 = p_i - p_i^2$$

Then the trace of covariance matrix is

$$
\begin{aligned}
tr(Cov(\mathbf{Y})_{)} &= \sum_{i=1}^{k} Cov(\mathbf{Y})_{ii} \\
&= \sum_{i=1}^{k} p_i - p_i^2 \\
&= 1 - \sum_{i=1}^{k} p_i^2
\end{aligned}
$$

(c) Above result measures purity of a split in a k-class decision tree. In other word, it is the expected error rate if the prediction is randomly selected based on the class distribution at that node.
For example at one split, the result region has p_i data belonging to class i. Then for an arbitrary sample that are drawn from the same empirical distribution, and we randomly label it, the probability of labeling correctly is

$$\sum_{i=1}^{k} p_i^2$$

Therefore the expected misclassification rate is

$$1 - \sum_{i=1}^{k} p_i^2$$