Recitation 9
Ensemble Methods, SVM, and Kernels

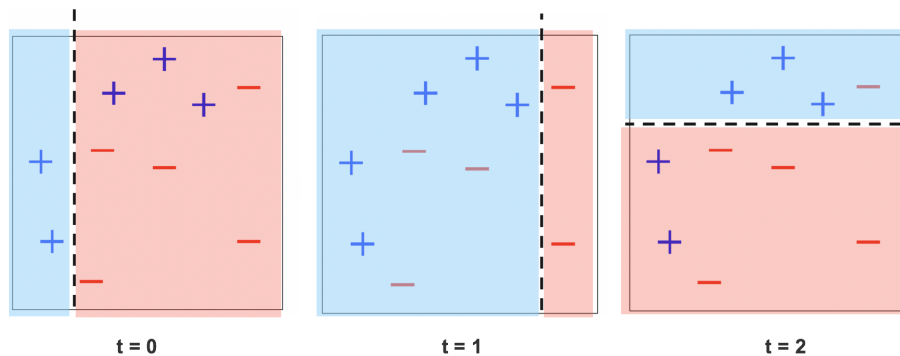10-701: Introduction to Machine Learning

11/20/2025

1 Ensemble Methods

The idea of ensemble methods is to build a model for prediction by combining the strengths of a group of simpler models. We'll cover two examples of ensemble methods: random forests and AdaBoost.
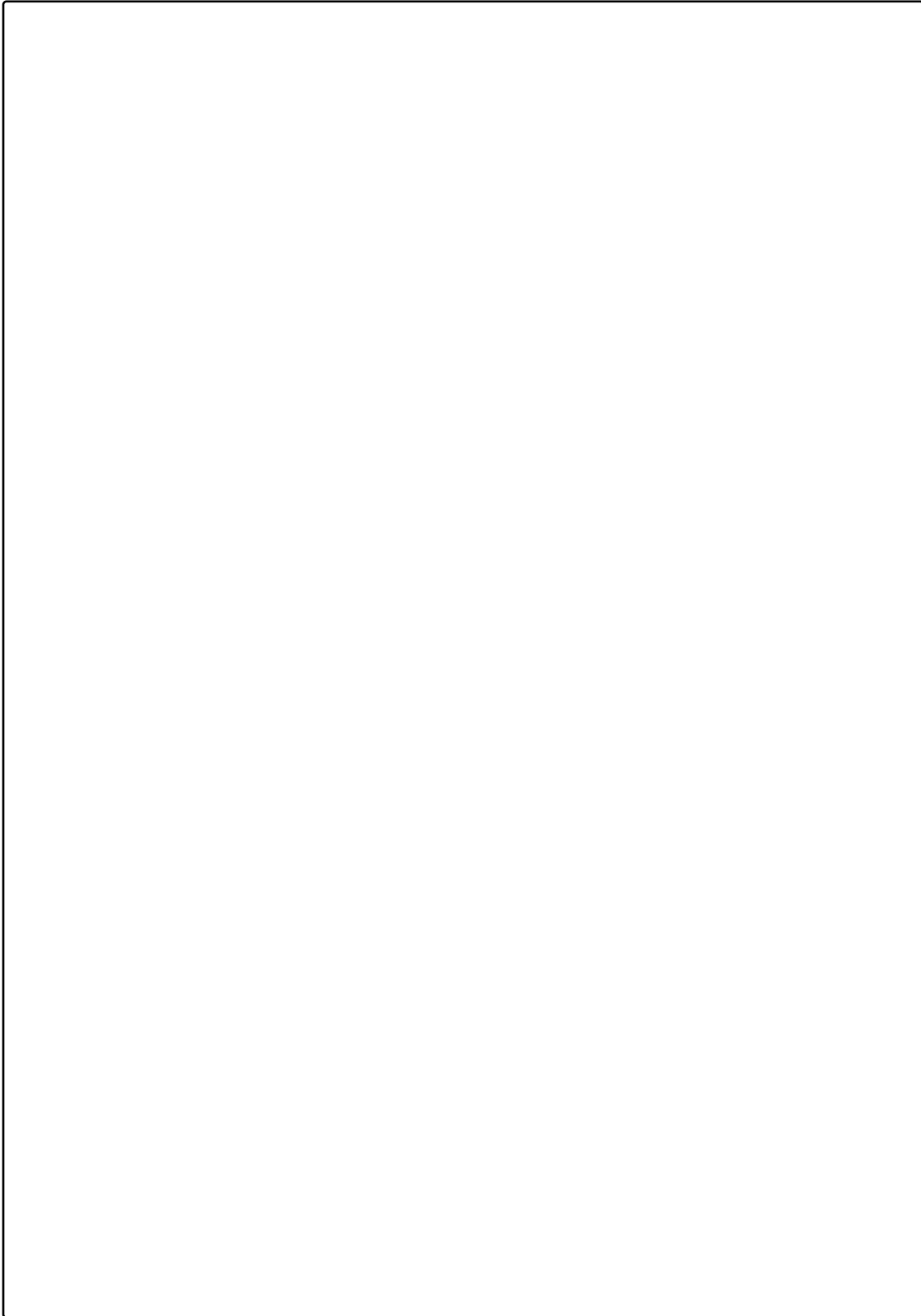
1.1 Adaboost

1.1.1 Practical Example

(Adapted from Eric Xing's 10701 slides) The graphs below show three iterations running Adaboost with a depth 1 decision tree. Each dashed line represents the decision boundary of h_t, and the shaded regions represent the predictons, positive (blue) or negative (red). For each iteration find the weighted training error ϵ_t and importance α_t of h_t. For $t = 0$ and $t = 1$ also find the weight normalization Z_t and record the updated weight for each point.

t = 0   t = 1   t = 2
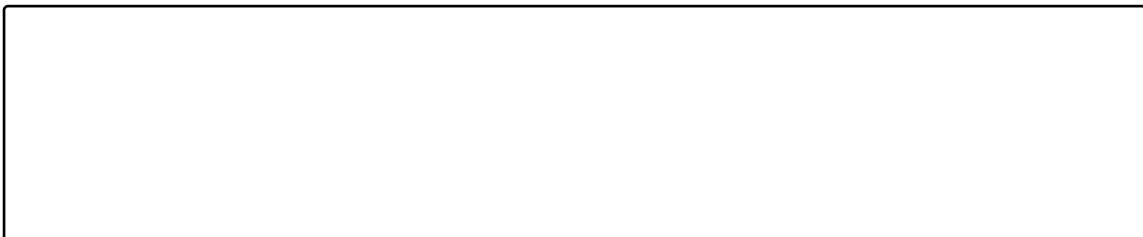
1. $\mathbf{t = 0}$

2. **t = 1**

3. **t = 2**

1.2   Random Forests

1. What are some downsides of decision trees, and how can we explain this in the context of the bias-variance tradeoff?

<div align="center">

Random Forests = Sample Bagging + Split-Feature Randomization

</div>

2. What is **sample bagging**?

3. What is **split-feature randomization**?

4. How do these techniques affect the bias and variance of an individual tree?

5. How do these techniques affect the bias and variance of an ensemble of trees?

6. For each data point $\mathbf{x}^{(i)}$, define $t^{(-i)}$ to be the set of decision trees that $\mathbf{x}^{(i)}$ was not used to train. Use each tree in $t^{(-i)}$ to make a prediction for $\mathbf{x}^{(i)}$, and use these predictions to make an aggregated prediction $\overline{t^{(-i)}}(\mathbf{x}^{(i)})$ (i.e. for classification take the majority vote). Then, we can define the *out-of-bag* error as follows:

$$E_{OOB} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left(\overline{t^{(-i)}}(\mathbf{x}^{(i)}) \neq y^{(i)}\right)$$

Why can we use E_{OOB} for hyperparameter optimization even though it was calculated using training points we used to learn the decision trees with?

7. **Random Forest Example:** Suppose we train a random forest with two decision trees on the following dataset, using the provided bootstrap samples. Assume that for ties, we predict $Y = 1$.

All	X_0	X_1	X_2	X_3	Y
1	1	0	0	0	1
2	0	0	1	0	1
3	0	0	0	1	1
4	0	0	0	0	0
5	0	1	0	1	1

Sample 1	X_0	X_1	X_2	X_3	Y
1	1	0	0	0	1
4	0	0	0	0	0
5	0	1	0	1	1

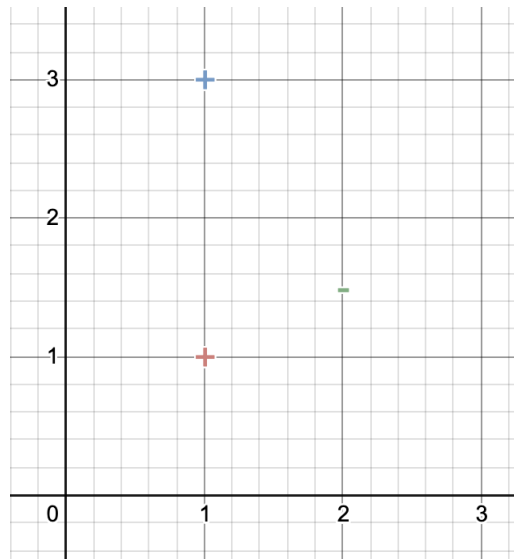| Sample 2	X_0	X_1	X_2	X_3	Y
3	0	0	0	1	1
4	0	0	0	0	0
5	0	1	0	1	1

(a) Suppose we train our first tree on Sample 1 and the split feature randomization chooses $\{X_1, X_2\}$ for the feature candidates at the root. What feature will we split on at the root?

(b) Suppose we then recurse on the left child (with feature value 0) of the root and split feature randomization chooses $\{X_0, X_2\}$ for the feature indices. What feature will we split on?

(c) Suppose we train our second tree on Sample 2 and the split feature randomization chooses $\{X_2, X_3\}$ for the feature candidates at the root. What feature will we split on at the root?

(d) What is the training error of the ensemble?

(e) What is the out of bag error of the ensemble?

2 SVM and Kernels

2.1 Questions on SVM with hard-margin

1. What is the decision boundary and the margin if we run a Hard-Margin SVM on the following set of points?

2. A few additional data points are added to the data set in Figures 3 (a) and 3 (b). Draw the new decision boundaries and give the margins corresponding to this boundaries. In which case does the decision boundary undergo a change and why?
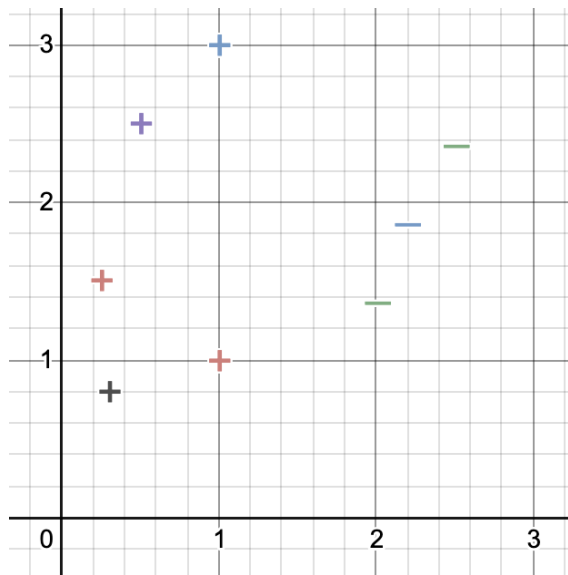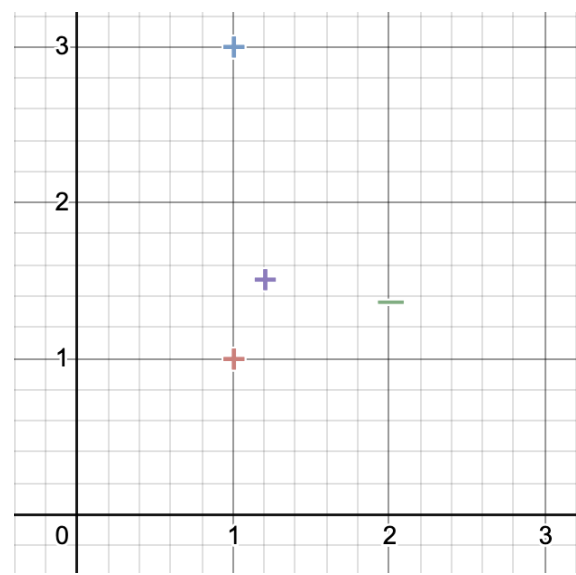
Figure 3(a)

Figure 3(b)

2.2 Soft-margin linear SVM

Given the following dataset in 1-d space (Figure 5), which consists of 4 positive data points and 3 negative data points. Suppose that we want to learn a soft-margin linear SVM for this data set. Remember that the soft-margin linear SVM can be formalized as the following constrained quadratic optimization problem. In this formulation, C is the regularization parameter, which balances the size of margin (i.e., smaller $w^T w$) vs. the violation of the margin (i.e. smaller $\sum_{i=1}^{m} \epsilon_i$).

$$\text{argmin}_{\{w,b\}} \frac{1}{2} w^t w + C \sum_{i=1}^{m} \epsilon_i$$
$$\text{Subject to} : y_i(w^t x_i + b) \geq 1 - \epsilon_i$$
$$\epsilon_i \geq 0 \ \forall i$$

Figure 5

1. If $C = \infty$, which means that we only care the violation of the margin, how many support vectors do we have?

2. If $C = 0$, which means that we only care about the size of the margin, how many support vectors do we have?

2.3   Margin trivia

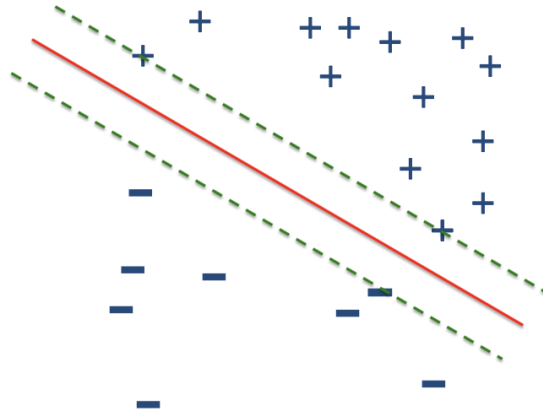SVMs try to find the linear separator that maximises the *margin* between datapoints.

Figure 1: SVM decision boundary for some sample data

1. How can we show that w is perpendicular to the decision boundary of SVMs?

2.4   Composite kernels

We can construct kernels by combining existing kernels in valid ways. We will verify three kernels of these properties by computing the implied feature transformations.

For any valid kernels K_1 and K_2 with implied feature transformations Φ_1 and Phi_2 and for nonnegative coefficients c_1, c_2, show that the following expressions are valid kernels:

1. $K(x, x') = c_1 K_1(x, x') + c_2 K_2(x, x')$

2. $K(x, x') = c_1 K_1(x, x') K_2(x, x')$

3. $K(x, x') = e^{K_1(x,x')}$