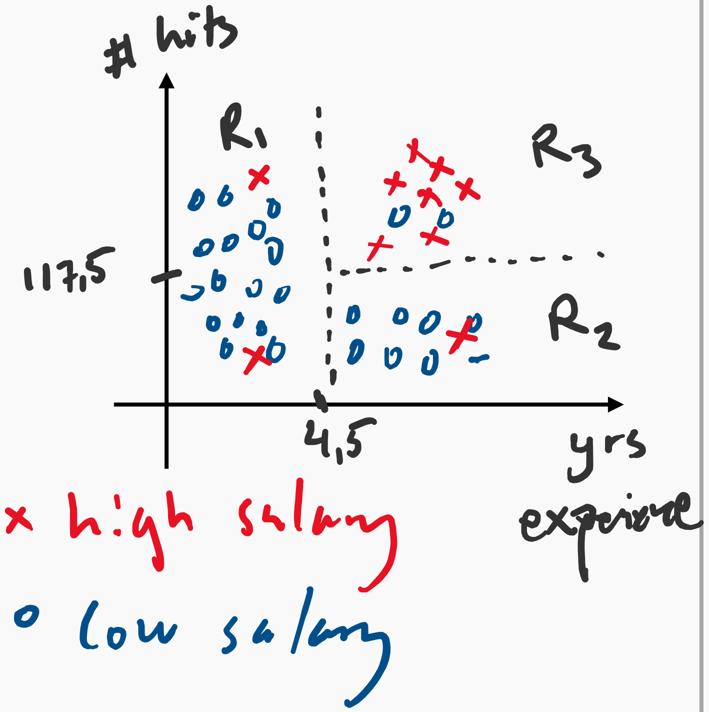
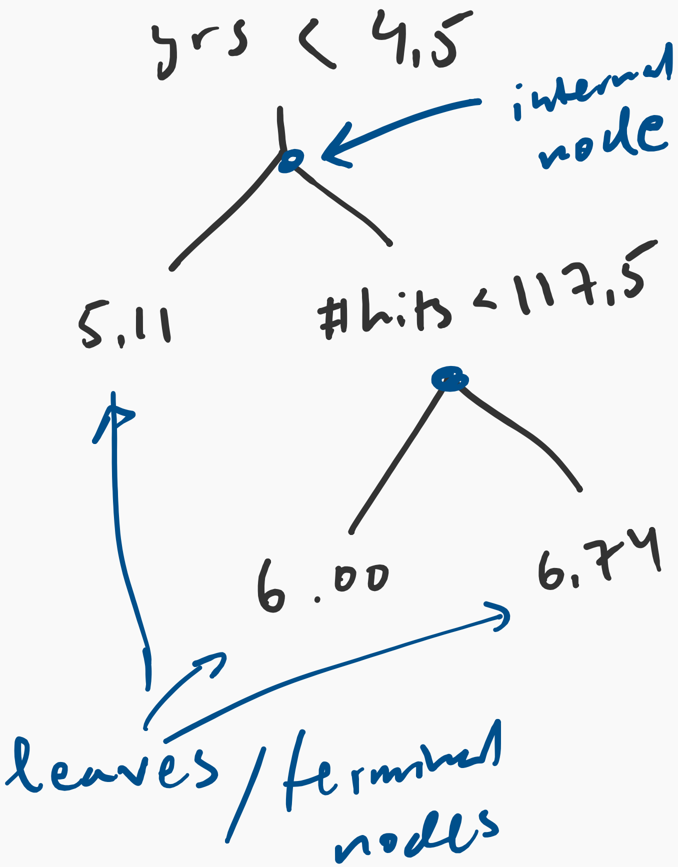
kokchun giang

growing a **decision tree** for classification and regression

grow a **tree** from the data

baseball players number
of hits and total years
experience





for regression

→ values of leaves are mean of all observations in region R_i

Interpretation

→ yrs affect salary most

→ ↓ experience → #hits don't affect salary as much

the process of building a **tree**

we've divided feature space into regions $R_1, \to R_J$

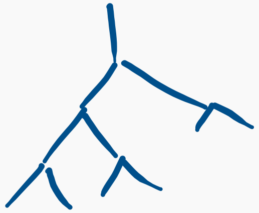
goal: find regions $R_1, \to R_J$ such that RSS is minimized

↑ a loss fcn for regression

Tree building process

1. recursive binary splitting → greedy
   → find the predictor & cutoff value to minimize RSS @ given step

2. repeat 1. until stopping criteria met



e.g. max 5 observations in each regions

→ overfitting

3. cost complexity pruning
   → grow large tree
   → prune to get subtrees
   → pick tree through cross-validat of α

classification tree

similar to regression tree but use gini index G instead of RSS in recursive binary splitting

G measures a nodes purity

↓ $G \Rightarrow$ ↑ purity

$\Rightarrow$ ↑ observations belong to one class