Federated Learning on Non-IID Data

分支主题 3

分支主题 2

Categories of Non-IID Data

Data tend to deviate from being identically distributed, that is $P_i = P_j$ for different clients i and j. Rewriting $P_i(x, y)$ as $P_i(y|x)P_i(x)$ and $P_i(x|y)P_i(y)$ allows us to characterize the differences more precisely.

Attribute skew

- Non-overlapping attribute skew
- Partial overlapping attribute skew
- Full overlapping attribute skew
  - Feature distribution skew (covariate shift)
    - Feature distribution skew (covariate shift): The marginal distributions Pi(x) may vary across clients, even if $P(y|x)$ is shared. For example, in a handwriting recognition domain, users who write the same words might still have different stroke width, slant, etc.
  - Same label, different features (concept drift)
    - The conditional distributions $P_i(x|y)$ may vary across clients even if $P(y)$ is shared. The same label y can have very different features x for different clients, e.g. due to cultural differences, weather effects, standards of living, etc. For example, images of homes can vary dramatically around the world and items of clothing vary widely. Even within the U.S., images of parked cars in the winter will be snow-covered only in certain parts of the country. The same label can also look very different at different times, and at different time scales: day vs. night, seasonal effects, natural disasters, fashion and design trends, etc.

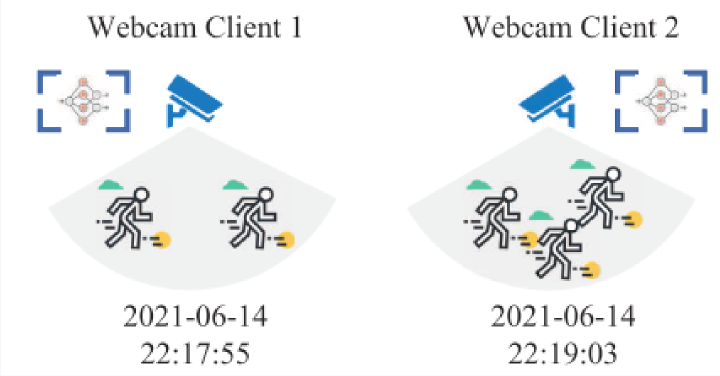
Label skew

- Label distribution skew (prior probability shift)
  - The marginal distributions $P_i(y)$ may vary across clients, even if $P(x|y)$ is the same. For example, when clients are tied to particular geo-regions, the distribution of labels varies across clients — kangaroos are only in Australia or zoos; a person's face is only in a few locations worldwide; for mobile device keyboards, certain emoji are used by one demographic but not others.
- Label preference skew / Same features, different label (concept shift)
  - The conditional distribution $P_i(y|x)$ may vary across clients, even if $P(x)$ is the same. Because of personal preferences, the same feature vectors in a training data item can have different labels. For example, labels that reflect sentiment or next word predictors have personal and regional variation.

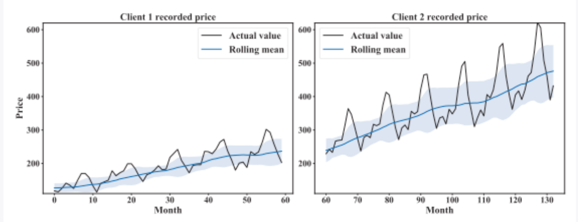
Temporal skew

Different clients hold data with different labels and different features, which integrates the characteristics of both horizontal and vertical FL.

- Spatio-temporal data

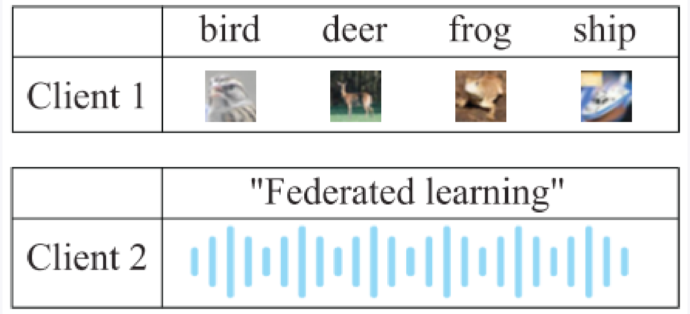


- Time-series data (Time-stamped data)



Other scenarios

- Attribute & Label skew



- Quantity skew
  - Different clients can hold vastly different amounts of data.