Purah P2 Lightning Talk

Ankit Lalwani, Shihan Ma, Nathan Wynn

Ethics - Key Takeaways

Collection

Large upfront cost, large payoff in the modeling stage.

YFCC100m pros and cons: diversity and skew.

Creative Commons license!

Data Storage

Data retention plans were unnecessary.

Analysis

Higher diversity -> lower bias towards physical characteristics.

Higher bias towards online Flickr community.

Modeling

Transparency.

Only using facial anatomic features, no personal information, no categorical information.

Deployment

Full control of rollbacks, no feature drift.

Educational purposes only!

Python Packages

- Pandas
  - Read CSV files
  - Select useful features (39 features)

- Scikit-learn
  - K-th nearest neighbors (KNN)
  - Linear Discriminative Analysis (LDA)
  - Quadratic Discriminant Analysis (QDA)

ML models

- Classification
  - K-th Nearest Neighbors (second best accuracy)
  - Linear Discriminative Analysis (best accuracy)
  - Quadratic Discriminant Analysis
  - Random Forest
  - Logistic Regression

- dimensionality reduction
  - Principal Component Analysis
  - Elastic Net Regression (with cross-validation)
    - Selects input variables with high coefficients.

Thank you!

Questions?