Recap and Review

Applied Machine Learning
Derek Hoiem

Midterm Exam Logistics

- Thurs, Mar 7 (start exam between 9:30 AM and 10:30 PM)
- Exam will be 75 minutes long (or longer for those with DRES accommodations)
- Mainly multiple choice / multiple select
  - No coding or complex calculations; mainly tests conceptual understanding
- You take it at home (open book) on PrairieLearn
- **Not cheating**
  - Consult notes, practice questions/answers, slides, internet, etc.
- **Cheating**
  - Talking to a classmate about the exam after one (but not both) of you has taken it
  - Getting help from another person during the exam
  - Obtaining past exam questions/answers
- You will not have time to look up all the answers, so do prepare by reviewing slides, lectures, AML book, and practice questions

Midterm Exam Central Topics

- How does **train/test error** depend on
  - Number of training samples
  - Complexity of model
- **Bias-variance trade-off**, including meaning of “bias” and “variance” for ML models and “overfitting”
- **Basic function/form/assumptions of classification/regression** models (KNN, NB, linear/logistic regression, trees, SVMs, boosted trees, random forests, ensembles)
- **Entropy/Information gain**
- Data organization and transformation: **clustering, PCA**
- **Latent variables and robustness**: EM, density estimation, robust estimation and fittting
- Gradient descent, **SGD**

KNN Usage Example: Deep Face

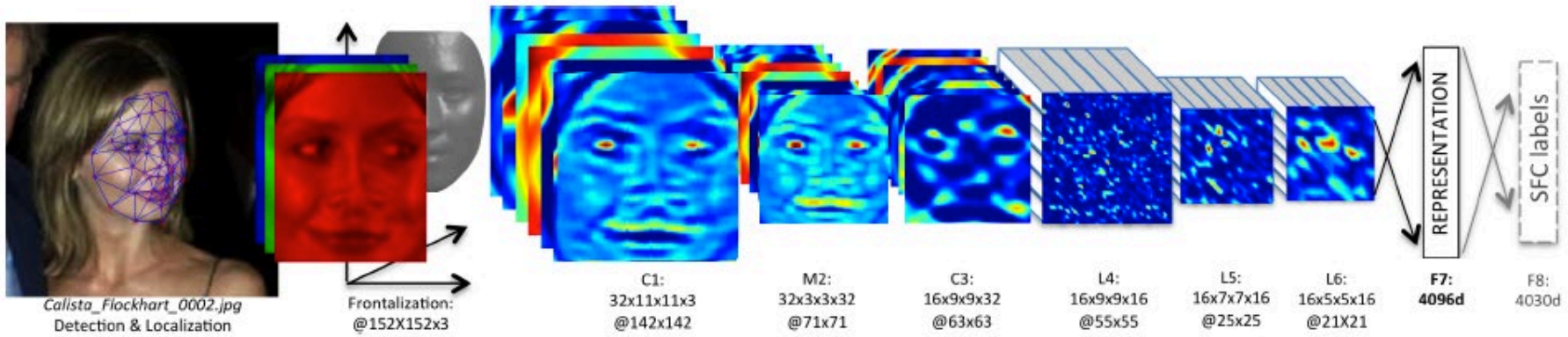
DeepFace: Closing the Gap to Human-Level Performance in Face Verification

Yaniv Taigman Ming Yang Marc'Aurelio Ranzato Lior Wolf

Facebook AI Research
Menlo Park, CA, USA
{yaniv, mingyang, ranzato}@fb.com

Tel Aviv University
Tel Aviv, Israel
wolf@cs.tau.ac.il

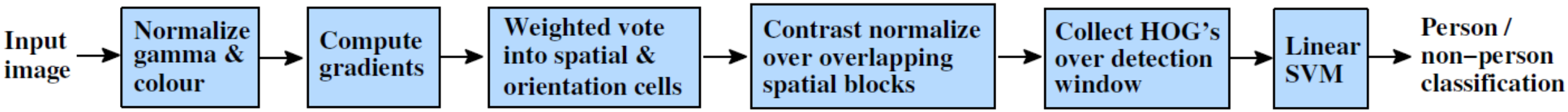
CVPR 2014



1. Detect facial features
2. Align faces to be frontal
3. Extract features using deep network while training classifier to label image into person (dataset based on employee faces)
4. In testing, extract features from deep network and use nearest neighbor classifier to assign identity

- Performs similarly to humans in the LFW dataset (labeled faces in the wild)
- Can be used to organize photo albums, identifying celebrities, or alert user when someone posts an image of them
- If this is used in a commercial deployment, what might be some unintended consequences?
- This algorithm is used by Facebook (though with expanded training data)

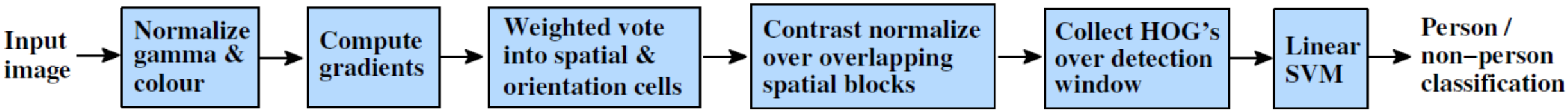
Example application of SVM: Dalal-Triggs 2005



- Detection by scanning window
  - Resize image to multiple scales and extract overlapping windows
  - Classify each window as positive or negative
- Very highly cited (40,000+) paper, mainly for HOG
- One of the best pedestrian detectors for several years

https://lear.inrialpes.fr/people/triggs/pubs/Dalal-cvpr05.pdf

Example application of SVM: Dalal-Triggs 2005



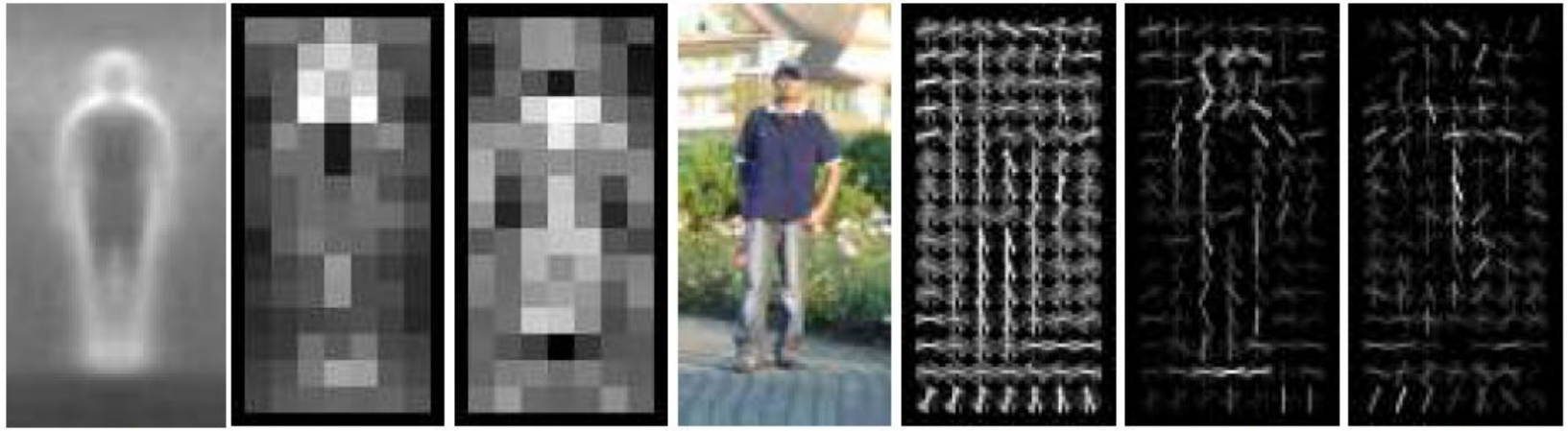
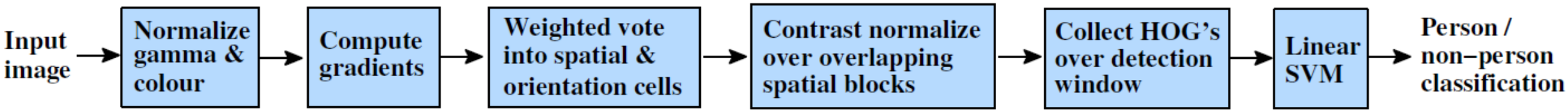
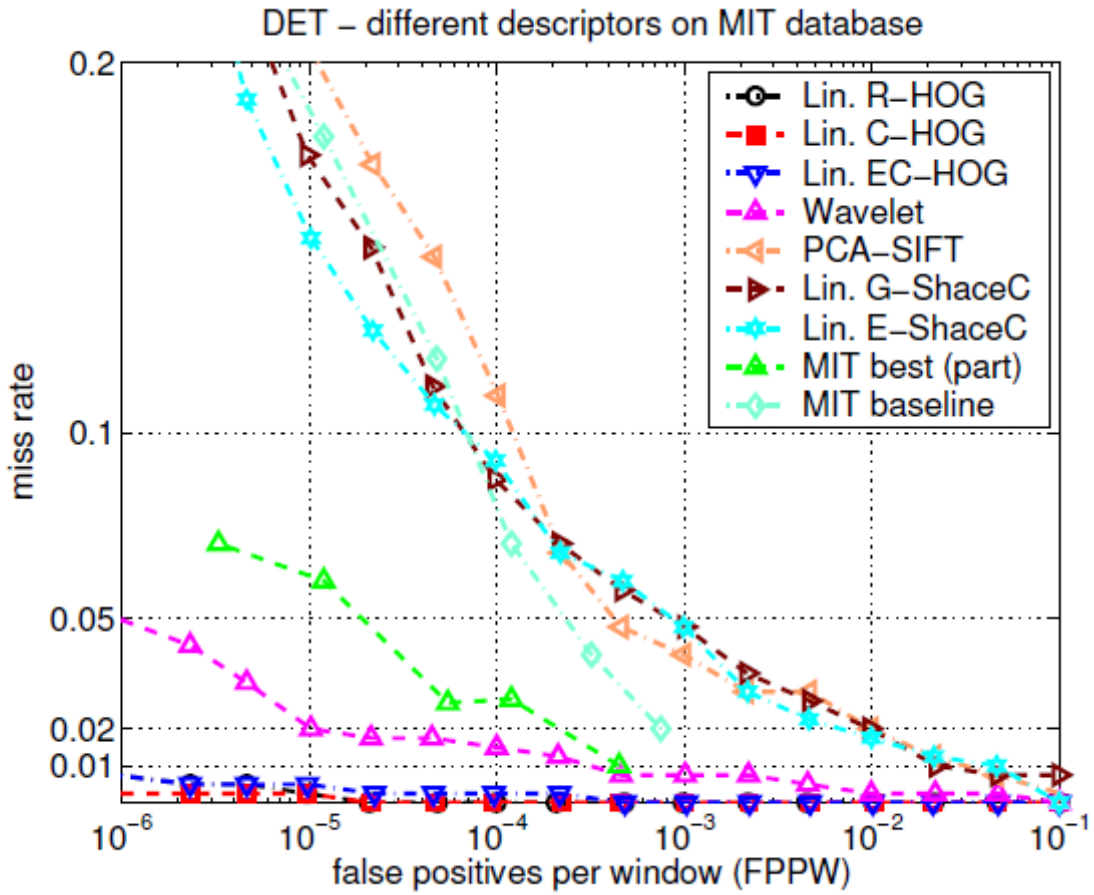


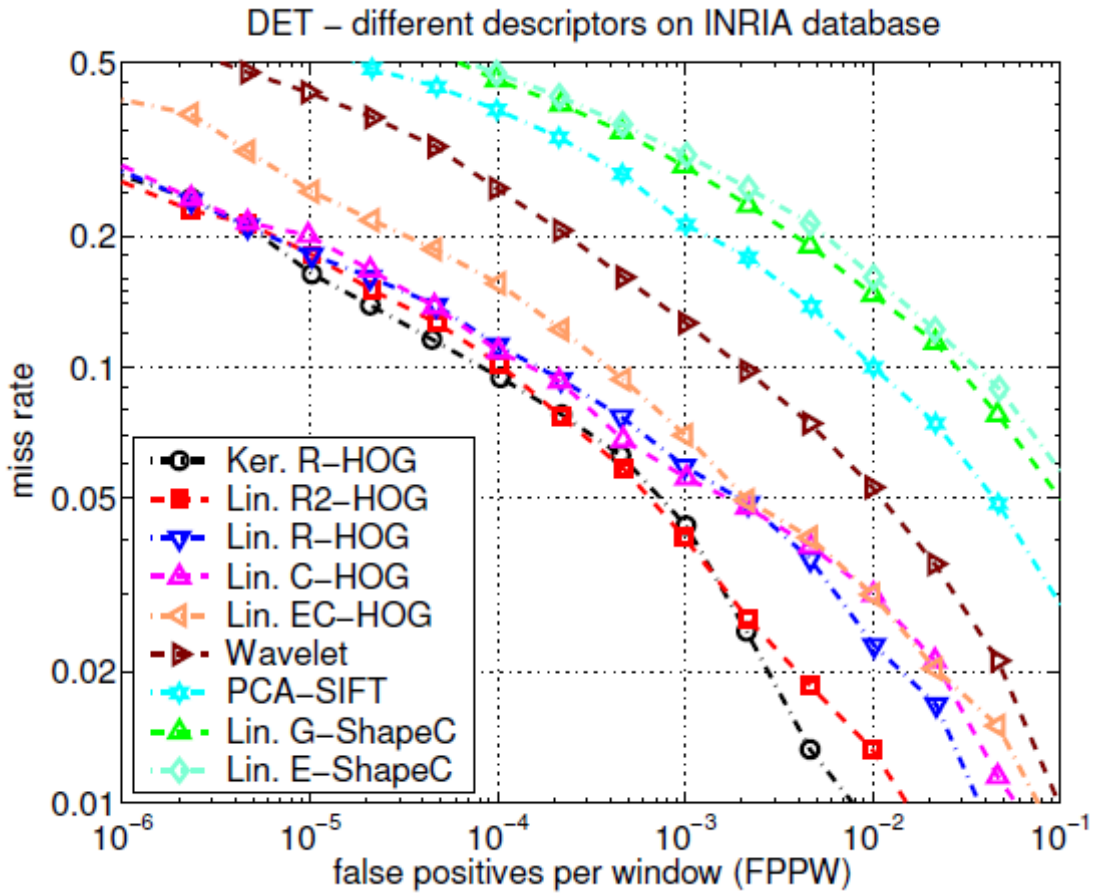
Figure 6. Our HOG detectors cue mainly on silhouette contours (especially the head, shoulders and feet). The most active blocks are centred on the image background just *outside* the contour. (a) The average gradient image over the training examples. (b) Each "pixel" shows the maximum positive SVM weight in the block centred on the pixel. (c) Likewise for the negative SVM weights. (d) A test image. (e) It's computed R-HOG descriptor. (f,g) The R-HOG descriptor weighted by respectively the positive and the negative SVM weights.

- Very highly cited (40,000+) paper, mainly for HOG
- One of the best pedestrian detectors for several years

Example application of SVM: Dalal-Triggs 2005





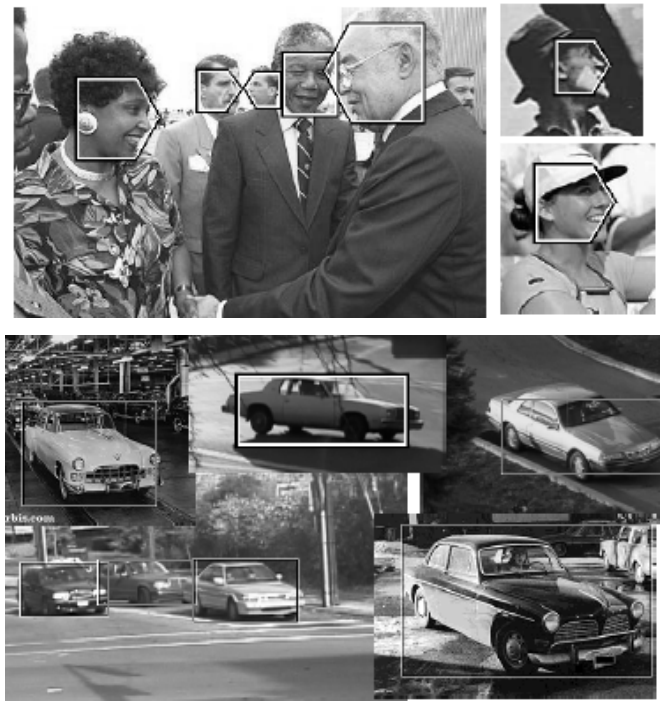


“Semi-naïve Bayes” object detection

- Best performing face/car detector in 2000-2005
- Model probabilities of small groups of features (wavelet coefficients)
- Search for groupings, discretize features, estimate parameters

A Statistical Method for 3D Object Detection Applied to Faces and Cars

Henry Schneiderman and Takeo Kanade

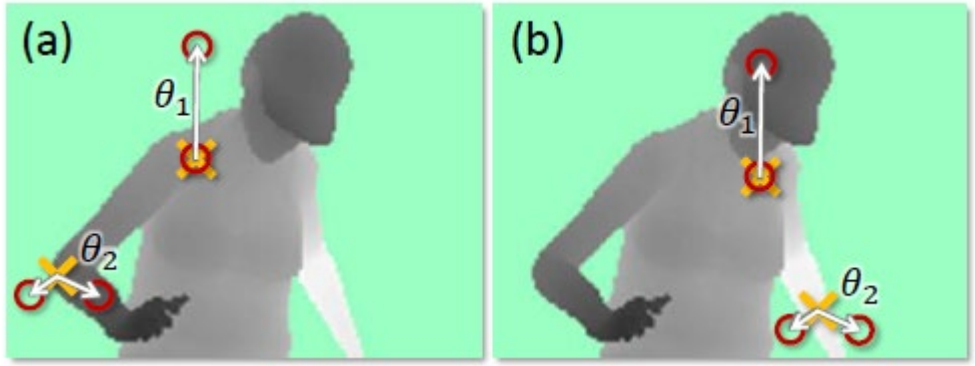


$$\frac{\prod_{x,y \in \text{region}} \prod_{k=1}^{17} P_k(pattern_k(x,y), x, y | \text{object})}{\prod_{x,y \in \text{region}} \prod_{k=1}^{17} P_k(pattern_k(x,y), x, y | \text{non-object})} > \lambda$$

https://www.cs.cmu.edu/afs/cs.cmu.edu/user/hws/www/CVPR00.pdf

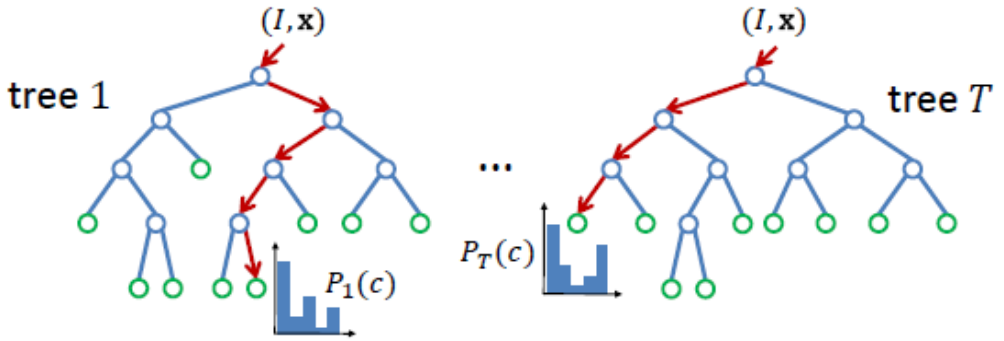
Human pose estimation with random forest

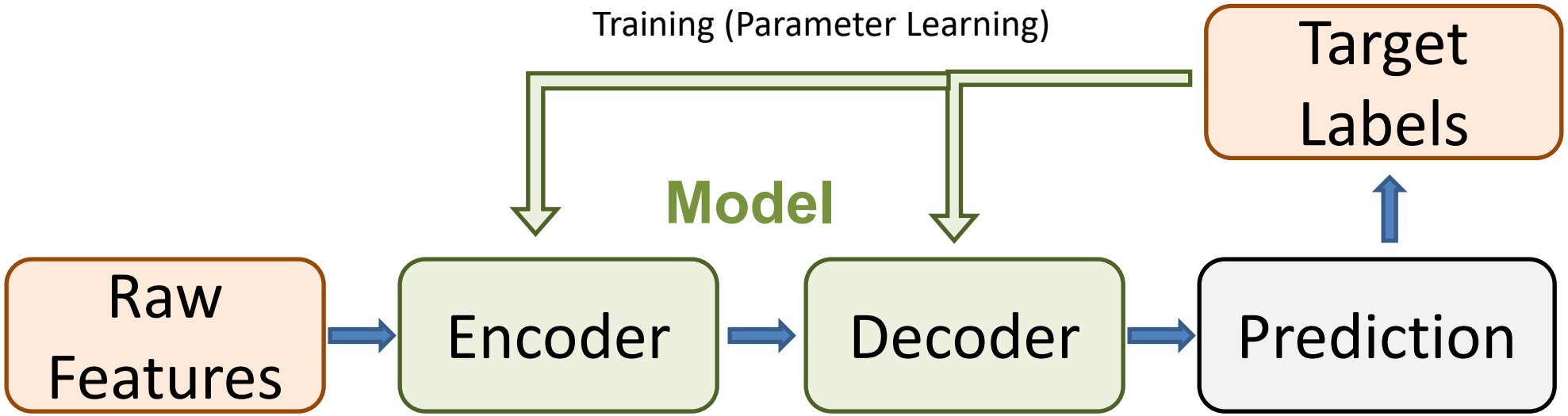
- Very simple features



- Lots of data

- Random Forest





Discrete/continuous values

Text
Images
Audio
Structured/unstructured
Few/many features
Clean/noisy labels

Trees
Feature selection
Clustering
Kernels
Density estimation

Manual feature design
Deep networks

Linear regressor
Logistic regressor
Nearest Neighbor
Probabilistic model
SVM

Category
Continuous value
Clusters
Low dimensional embedding

Pixel labels
Generated text, image, audio
Positions

Learning a model

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, Loss(f(\boldsymbol{X};\theta), \boldsymbol{y})$$

- $f(\boldsymbol{X};\theta)$: the model, e.g. $y = \boldsymbol{w}^T\boldsymbol{x}$
- θ: parameters of the model (e.g. $\boldsymbol{w}$)
- $(\boldsymbol{X}, \boldsymbol{y})$: pairs of training samples
- $Loss()$: defines what makes a good model
  - Good predictions, e.g. minimize $-\sum_n \log P(y_n|\boldsymbol{x}_n)$
  - Likely parameters, e.g. minimize $\boldsymbol{w}^T\boldsymbol{w}$
    - Regularization and priors indicate preference for particular solutions, which tends to improve generalization (for well chosen parameters) and can be necessary to obtain a unique solution

Prediction using a model

$$y_t = f(\boldsymbol{x_t}; \theta)$$

- Given some new set of input features $\boldsymbol{x}_t$, model predicts y_t
  - Regression: output y_t directly, possibly with some variance estimate
  - Classification
    - Output most likely y_t directly, as in nearest neighbor
    - Output $P(y_t|\boldsymbol{x}_t)$, as in logistic regression

Model evaluation process

1. Collect/define training, validation, and test sets
2. Decide on some candidate models and hyperparameters
3. For each candidate:
   a. Learn parameters with training set
   b. Evaluate trained model on the validation set
4. Select best model
5. Evaluate best model's performance on the test set
   - Cross-validation can be used as an alternative
   - Common measures include error or accuracy, root mean squared error, precision-recall

How to think about ML algorithms

- What is the model?
  - What kinds of functions can it represent?
  - What functions does it prefer? (regularization/prior)
- What is the objective function?
  - What “values” are implied?
  - The objective function does not always match the final evaluation metric
  - Objectives are designed to be optimizable and improve generalization
- How do I optimize the model?
  - How long does it take to train, and how does it depend on the amount of training data or number of features?
  - Can I reach a global optimum?
- How does the prediction work?
  - How accurate is the prediction?
  - How fast can I make a prediction for a new sample?
  - Does my algorithm provide a confidence on its prediction?

Bias-Variance Trade-off

$$\underbrace{E_{\mathbf{x},y,D}\left[(h_D(\mathbf{x}) - y)^2\right]}_{\text{Expected Test Error}} = \underbrace{E_{\mathbf{x},D}\left[(h_D(\mathbf{x}) - \bar{h}(\mathbf{x}))^2\right]}_{\text{Variance}} + \underbrace{E_{\mathbf{x},y}\left[(\bar{y}(\mathbf{x}) - y)^2\right]}_{\text{Noise}} + \underbrace{E_{\mathbf{x}}\left[(\bar{h}(\mathbf{x}) - \bar{y}(\mathbf{x}))^2\right]}_{\text{Bias}^2}$$

Variance: due to limited data
Different training samples will give different models that vary in predictions for the same test sample

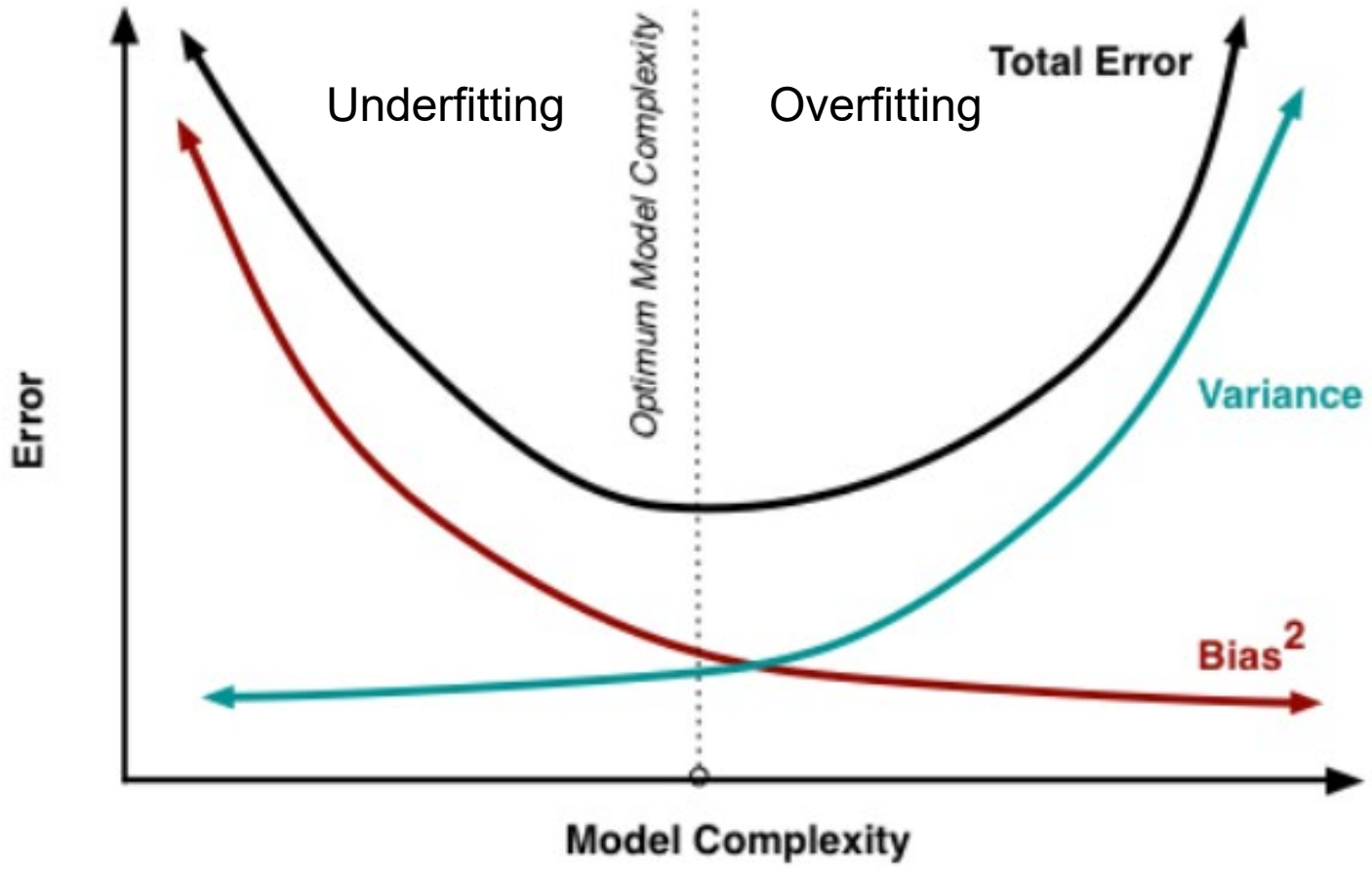
"**Noise**": irreducible error due to data/problem

Bias: error when optimal model is learned from infinite data

Above is for regression.
But same error = variance + noise + bias^2 holds for classification error and logistic regression.

See this for derivation

Fig Sources



How to detect high variance:

- Test error is much higher than training error

How to detect high bias or noise:

- The training error is high

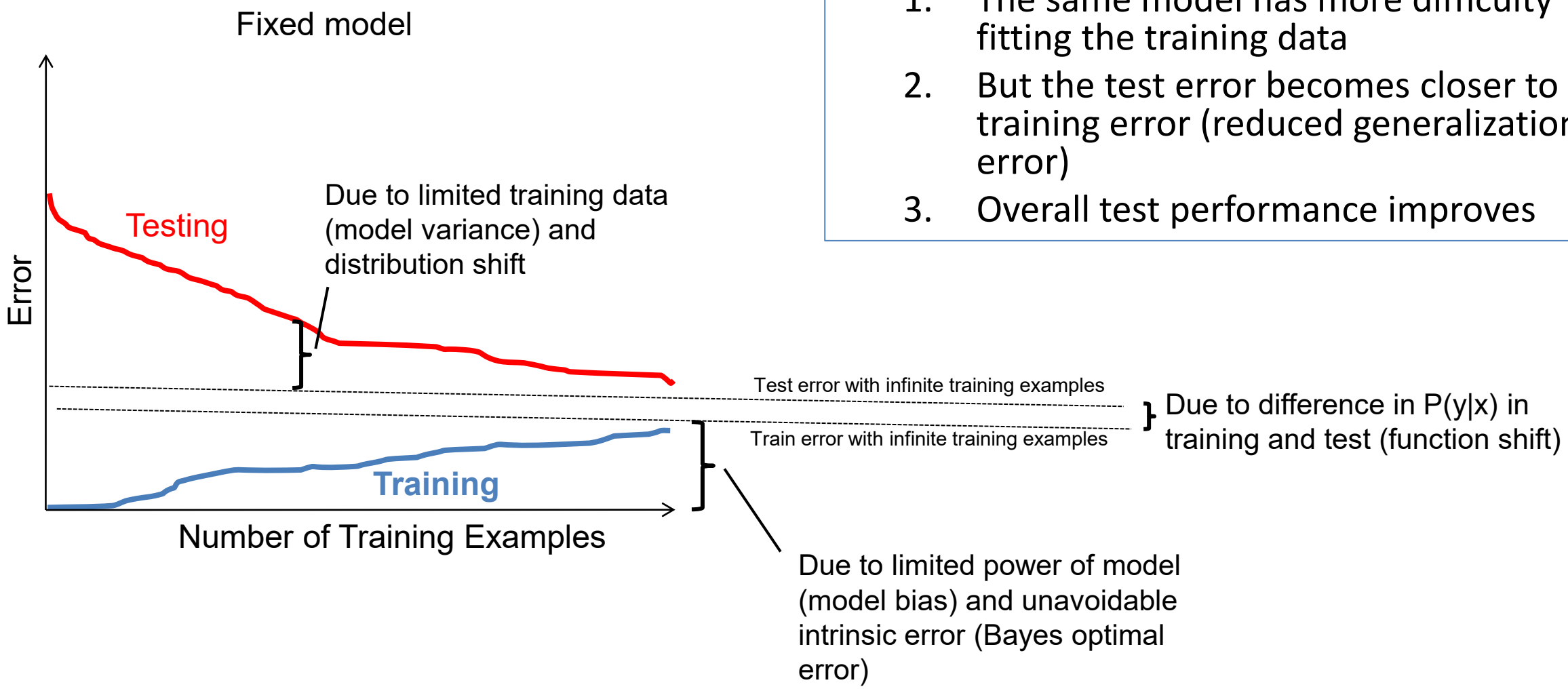
As you increase model complexity:

- Training error will decrease
- Test error may decrease (if you are currently “underfitting”) or increase (if you are “overfitting”)

What does “model complexity” mean?

- More parameters in the same structure, e.g. a deeper tree is more complex than a shallow tree
- Less regularization, e.g. smaller regularization penalty

Performance vs training size



As we get more training data:

1. The same model has more difficulty fitting the training data
2. But the test error becomes closer to training error (reduced generalization error)
3. Overall test performance improves

Classification methods

	Nearest Neighbor	Naïve Bayes	Logistic Regression	Decision Tree
Type	Instance-Based	Probabilistic	Probabilistic	Probabilistic
Decision Boundary	Partition by example distance	Usually linear	Usually linear	Partition by selected boundaries
Model / Prediction	$i^* = \underset{i}{\operatorname{argmin}}\, dist(X_{trn}[i], x)$ $y^* = y_{trn}[i^*]$	$y^* = \underset{y}{\operatorname{argmax}} \prod_i P(x_i \mid y) P(y)$	$w^T x + b \approx \log \frac{P(y=1 \mid x)}{P(y=0 \mid x)}$ $y^* = \underset{y}{\operatorname{argmax}}\, P(y \mid x)$	Conjunctive rules $y^* = leaf(x)$
Strengths	* Low bias * No training time * Widely applicable * Simple	* Estimate from limited data * Simple * Fast training/prediction	* Powerful in high dimensions * Widely applicable * Good confidence estimates * Fast prediction	* Explainable decision function * Widely applicable * Does not require feature scaling
Limitations	* Relies on good input features * Slow prediction (in basic implementation)	* Limited modeling power	* Relies on good input features	* One tree tends to either generalize poorly or underfit the data

Classification methods (extended)

assuming x in {0 1}

	Learning Objective	Training	Inference
Naïve Bayes	$\text{maximize} \sum_i \left[\sum_j \log P(x_{ij} \mid y_i; \theta_j) + \log P(y_i; \theta_0) \right]$	$\theta_{kj} = \frac{\sum_i \delta(x_{ij} = 1 \wedge y_i = k) + r}{\sum_i \delta(y_i = k) + Kr}$	$\theta_1^T \mathbf{x} + \theta_0^T (1 - \mathbf{x}) > 0$ where $\theta_{1j} = \log \frac{P(x_j = 1 \mid y = 1)}{P(x_j = 1 \mid y = 0)}$, $\theta_{0j} = \log \frac{P(x_j = 0 \mid y = 1)}{P(x_j = 0 \mid y = 0)}$
Logistic Regression	$\text{minimize} \sum_i -\log(P(y_i \mid \mathbf{x}, \boldsymbol{\theta})) + \lambda \lVert \boldsymbol{\theta} \rVert$ where $P(y_i \mid \mathbf{x}, \boldsymbol{\theta}) = 1/(1 + \exp(-y_i \boldsymbol{\theta}^T \mathbf{x}))$	Gradient descent	$\boldsymbol{\theta}^T \mathbf{x} > t$
Linear SVM	$\text{minimize } \lambda \sum_i \xi_i + \frac{1}{2} \lVert \boldsymbol{\theta} \rVert$ $\text{such that } y_i \boldsymbol{\theta}^T \mathbf{x} \geq 1 - \xi_i \ \ \forall i, \ \xi_i \geq 0$	Quadratic programming or subgradient opt.	$\boldsymbol{\theta}^T \mathbf{x} > t$
Kernelized SVM	complicated to write	Quadratic programming	$\sum_i y_i \alpha_i K(\hat{\mathbf{x}}_i, \mathbf{x}) > 0$
Nearest Neighbor	most similar features → same label	Record data	y_i $\text{where } i = \underset{i}{\text{argmin}} \ K(\hat{\mathbf{x}}_i, \mathbf{x})$

* Notation may differ from previous slide

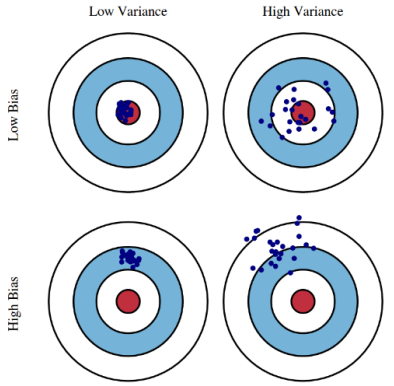
Regression methods

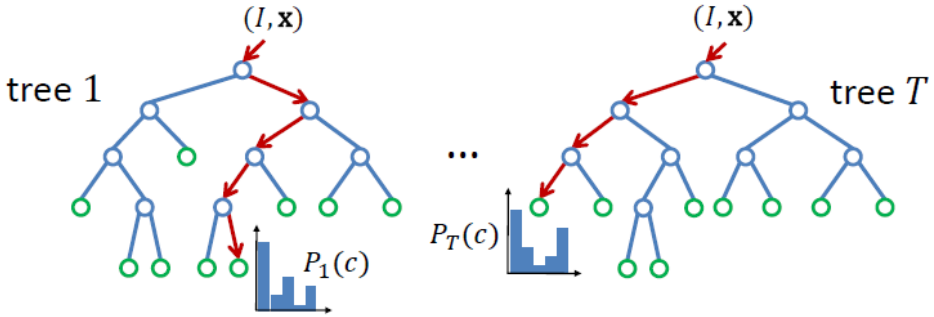
	Nearest Neighbor	Naïve Bayes	Linear Regression	Decision Tree
Type	Instance-Based	Probabilistic	Data fit	Probabilistic
Decision Boundary	Partition by example distance	Usually linear	Linear	Partition by selected boundaries
Model / Prediction	$i^* = \underset{i}{\text{argmin}}\, dist(X_{trn}[i], x)$ $y^* = y_{trn}[i^*]$	$y^* = \underset{y}{\text{argmax}} \prod_i P(x_i \mid y) P(y)$	$y^* = w^T x$	Conjunctive rules $y^* = leaf(x)$
Strengths	* Low bias * No training time * Widely applicable * Simple	* Estimate from limited data * Simple * Fast training/prediction	* Powerful in high dimensions * Widely applicable * Fast prediction * Coefficients may be interpretable	* Explainable decision function * Widely applicable * Does not require feature scaling
Limitations	* Relies on good input features * Slow prediction (in basic implementation)	* Limited modeling power	* Relies on good input features	* One tree tends to either generalize poorly or underfit the data

Ensembles

- Ensembles improve accuracy by reducing bias and/or variance
- Boosting minimizes bias by fixing previous mistakes, e.g. Boosted Decision Tree classifier
- Averaging over predictions from multiple models minimizes variance, e.g. Random Forests
- Random forests and boosted trees are powerful classifiers and useful for a wide variety of problems

$$\underbrace{E_{\mathbf{x},y,D}\left[(h_D(\mathbf{x}) - y)^2\right]}_{\text{Expected Test Error}} = \underbrace{E_{\mathbf{x},D}\left[(h_D(\mathbf{x}) - \bar{h}(\mathbf{x}))^2\right]}_{\text{Variance}} + \underbrace{E_{\mathbf{x},y}\left[(\bar{y}(\mathbf{x}) - y)^2\right]}_{\text{Noise}} + \underbrace{E_{\mathbf{x}}\left[(\bar{h}(\mathbf{x}) - \bar{y}(\mathbf{x}))^2\right]}_{\text{Bias}^2}$$





Questions

https://tinyurl.com/cs441midtermreview

Summaries

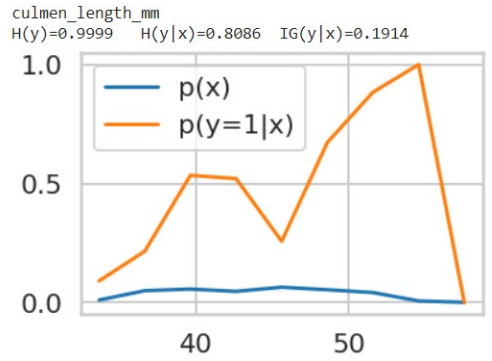
Working with Data (L2)

Machine learning is fitting parameters of a model so that you can accurately predict one set of numbers from another set of numbers

$$f(\boldsymbol{x};\theta) \rightarrow y$$

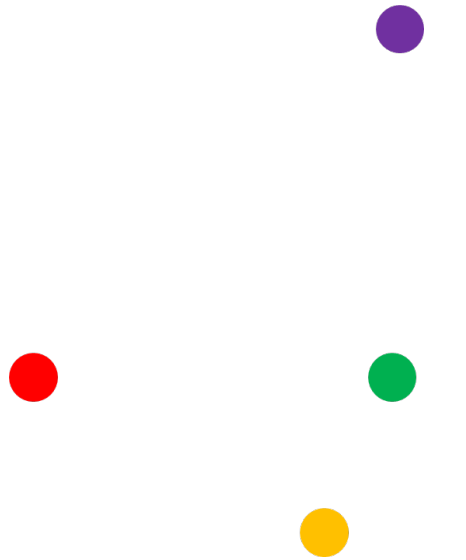
Something can take a lot of data storage but provide little information, or vice versa

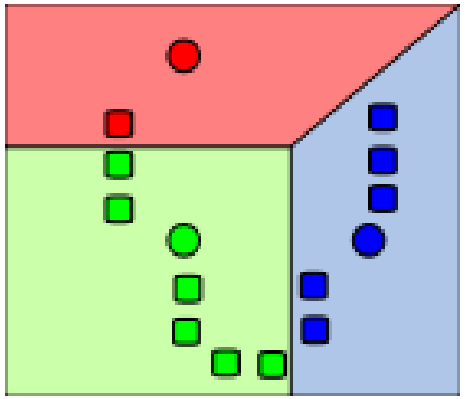
The predictiveness or information gain of the features depends on how they are modeled



Clustering (L3)

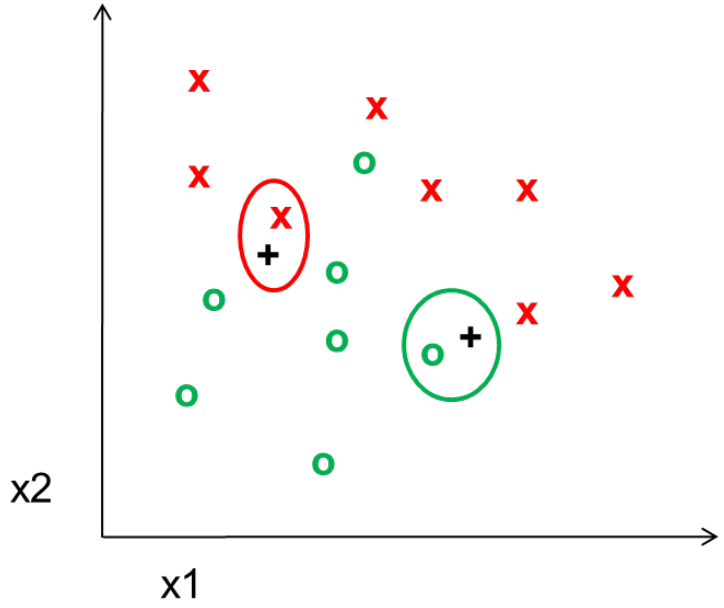
- Similarity is foundational to machine learning
- Use highly optimized libraries like FAISS for search/retrieval
- Approximate search methods like LSH can be used to find similar points quickly
- TF-IDF is used for similarity of tokenized documents and used with index for fast search
- Clustering groups similar data points
- K-means is the must-know method, but there are many others

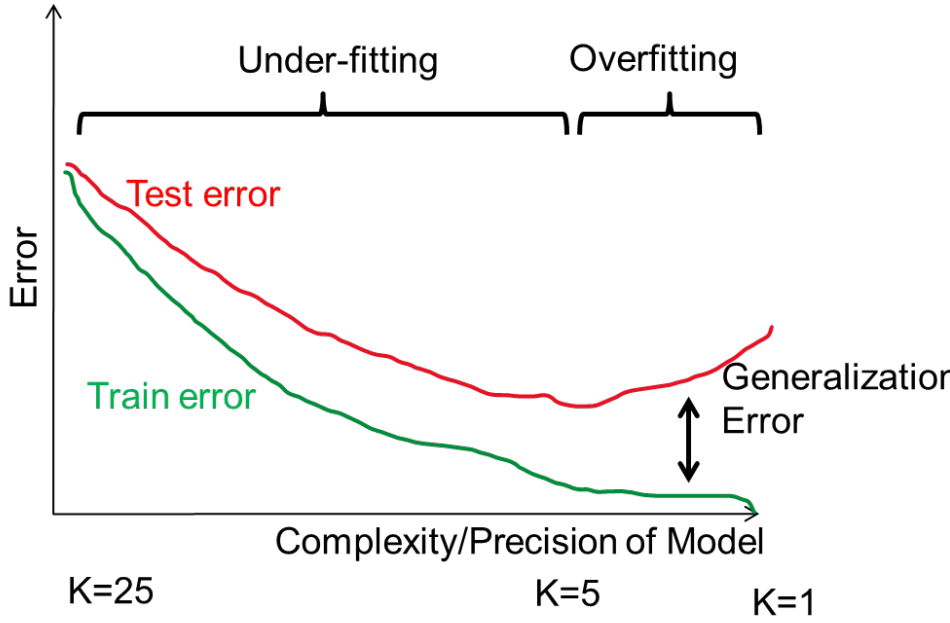




KNN (L4)

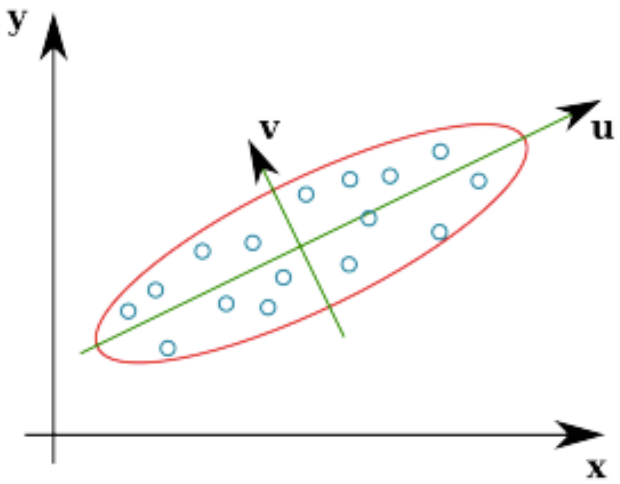
- KNN is a simple but effective classifier/regressor that predicts the label of the most similar training example(s)
- Larger K gives a smoother prediction function
- Test error is composed of bias (model too simple/smooth to fit data) and variance (model too complex to learn from training data)



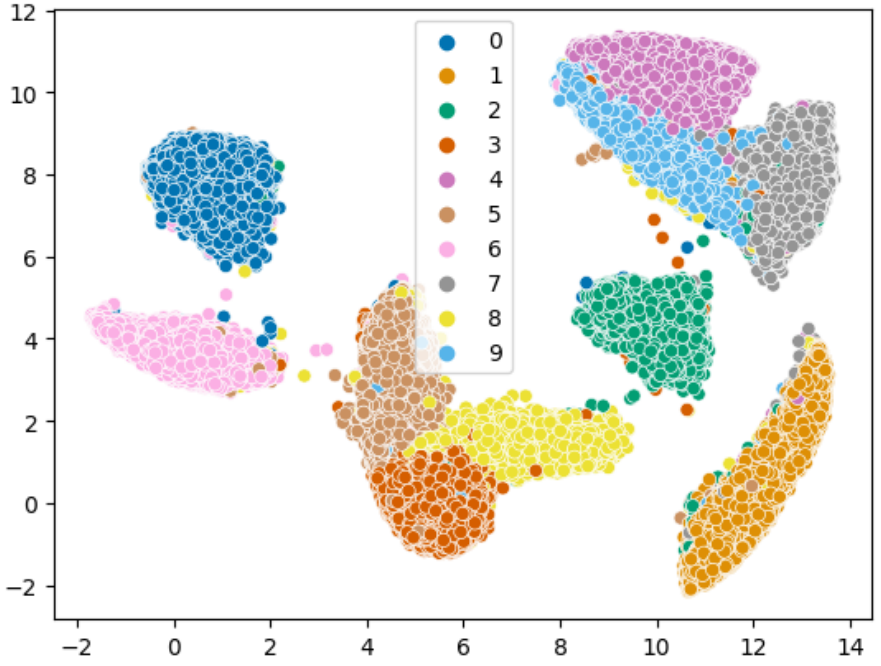


PCA/Embedding (L5)

- PCA reduces dimensions by linear projection
  - Preserves variance to reproduce data as well as possible, according to mean squared error
  - May not preserve local connectivity structure or discriminative information



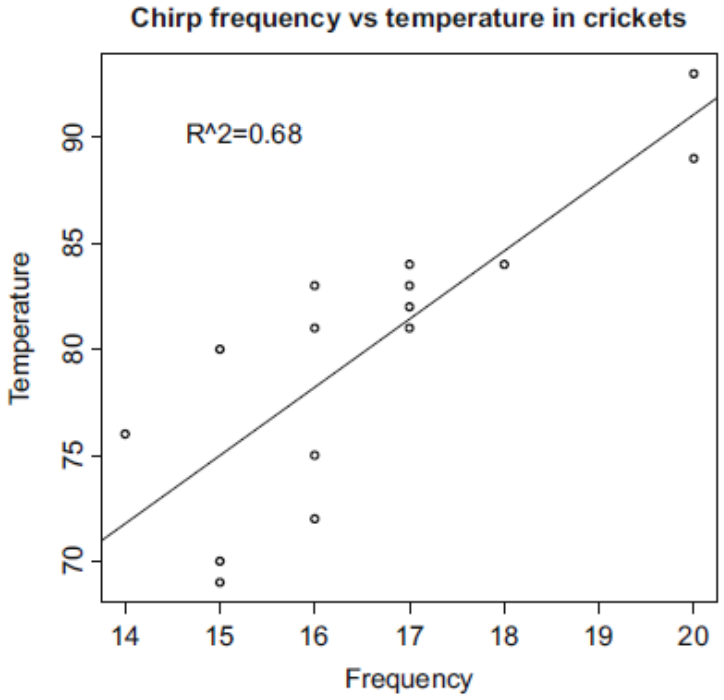
- Other methods try to preserve relationships between points
  - MDS: preserve pairwise distances
  - IsoMap: MDS but using a graph-based distance
  - t-SNE: preserve a probabilistic distribution of neighbors for each point (also focusing on closest points)
  - UMAP: incorporates k-nn structure, spectral embedding, and more to achieve good embeddings relatively quickly

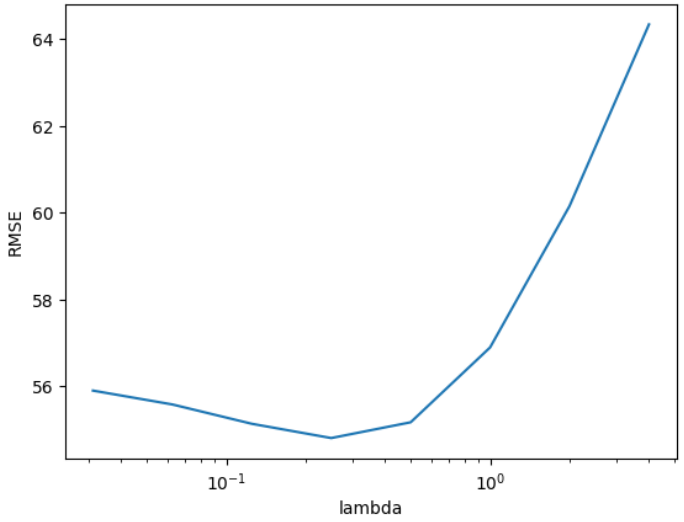


Linear Regression (L6)

- Linear regression fits a linear model to a set of feature points to predict a continuous value
  - Explain relationships
  - Predict values
  - Extrapolate observations

- Regularization prevents overfitting by restricting the magnitude of feature weights
  - L1: prefers to assign a lot of weight to the most useful features
  - L2: prefers to assign smaller weight to everything





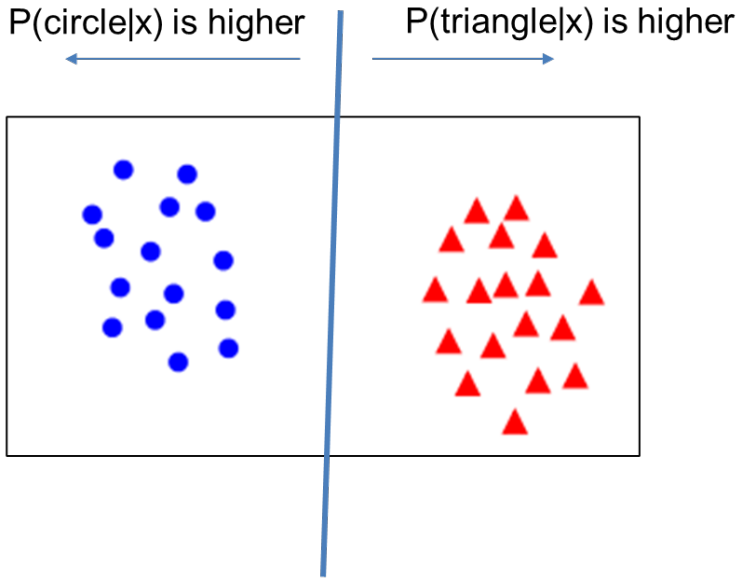
Linear Classifiers (L7)

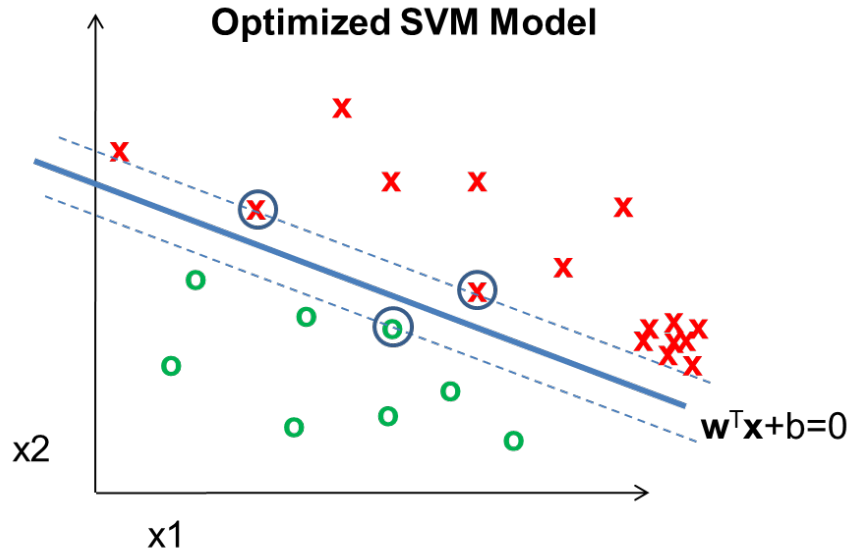
- Linear logistic regression and linear SVM are classification techniques that aims to split features between two classes with a linear model
  - Predict categorical values with confidence

- Logistic regression maximizes confidence in the correct label, while SVM just tries to be confident enough

- Non-linear versions of SVMs can also work well and were once popular (but almost entirely replaced by deep networks)

- Nearest neighbor and linear models are the final predictors of most ML algorithms – the complexity lies in finding features that work well with NN or linear models





Probability / Naïve Bayes (L8)

- Probabilistic models are a large class of machine learning methods

- Naïve Bayes assumes that features are independent given the label
  - Easy/fast to estimate parameters
  - Less risk of overfitting when data is limited

$$P(\boldsymbol{x}, y) = \prod_i P(x_i|y)P(y)$$

- You can look up how to estimate parameters for most common probability models
  - Or take partial derivative of total data/label likelihood given parameter

- Prediction involves finding y that maximizes $P(x, y)$, either by trying all y or solving partial derivative

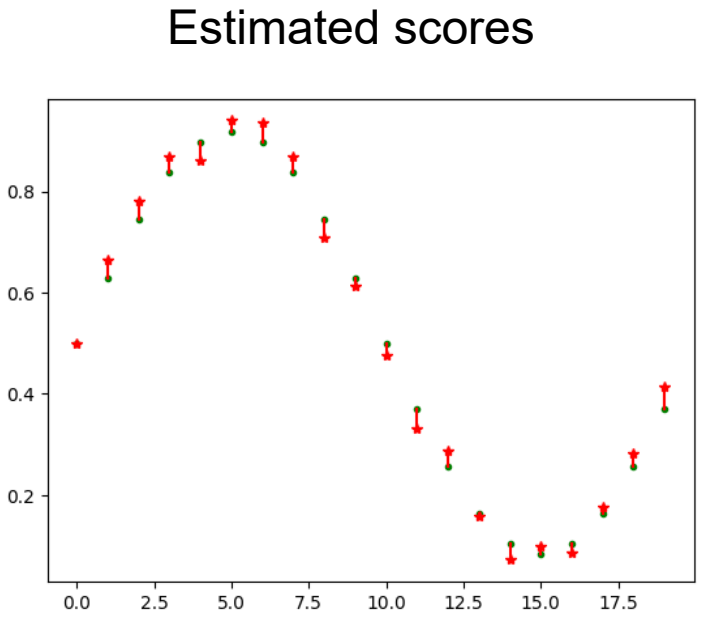
- Maximizing $\log P(x, y)$ is equivalent to maximizing $P(x, y)$ and often much easier

$$y^* = \operatorname{argmax}_y \prod_i P(x_i|y)\,P(y)$$

$$= \operatorname{argmax}_y \sum_i \log P(x_i|y) + \log P(y)$$

EM (L9)

- EM is a widely applicable algorithm to solve for latent variables and parameters that make the observed data likely
  - E-step: compute the likelihoods of the values of the latent variables
  - M-step: solve for most likely model parameters, using the likelihoods from the E-step as weights
- While derivation is long and somewhat complicated, the application is simple
- EM is used, for example, in mixture of Gaussian and topic models



Good annotators: 0, 1, 3

PDF Estimation (L10)

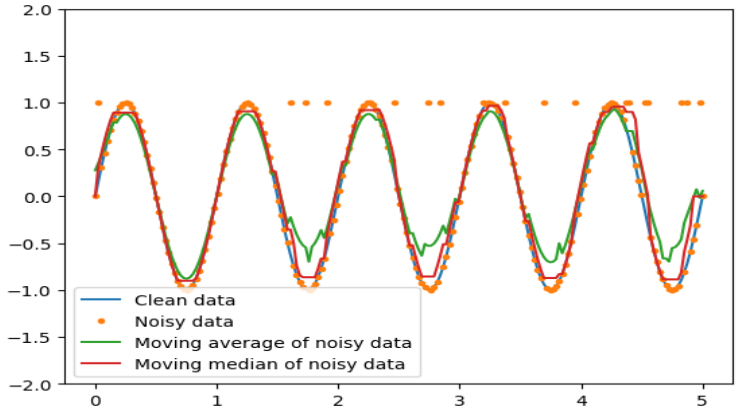
	Parametric Models	**Semi-Parametric**	**Non-Parametric**
Description	Assumes a fixed form for density	Can fit a broad range of functions with limited parameters	Can fit any distribution
Examples	Gaussian, exponential	Mixture of Gaussians	Discretization, kernel density estimation
Good when	Model is able to approximately fit the distribution	Low dimensional or smooth distribution	1-D data
Not good when	Model cannot approximate the distribution	Distribution is not smooth, challenging in high dimensions	Data is high dimensional

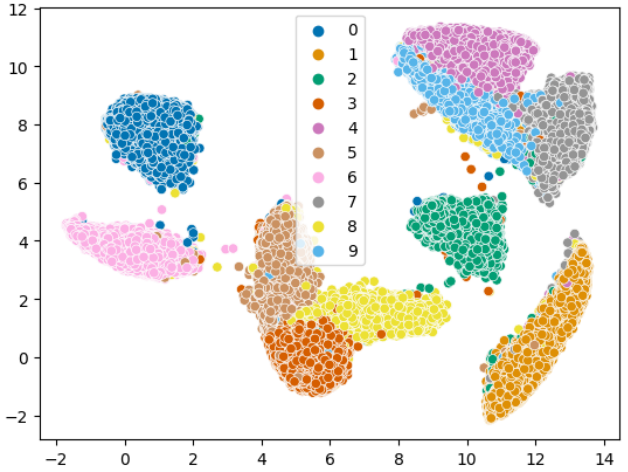
Robust Estimation (L11)

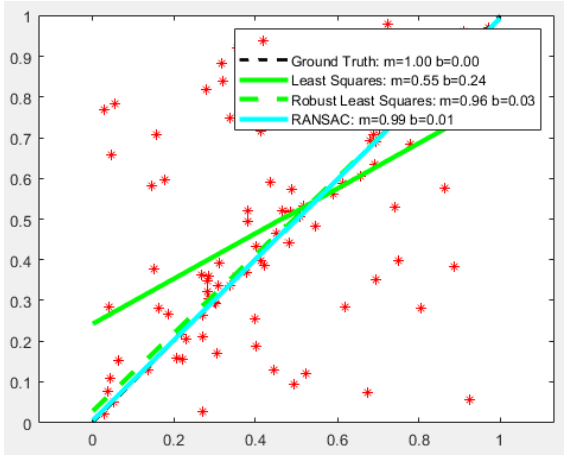
Median and quantiles are robust to outliers, while mean/min/max aren't

Outliers can be detected as low probability points, low density points, poorly compressible points, or through 2D visualizations

Least squares is not robust to outliers. Use RANSAC or IRLS or robust loss function instead.

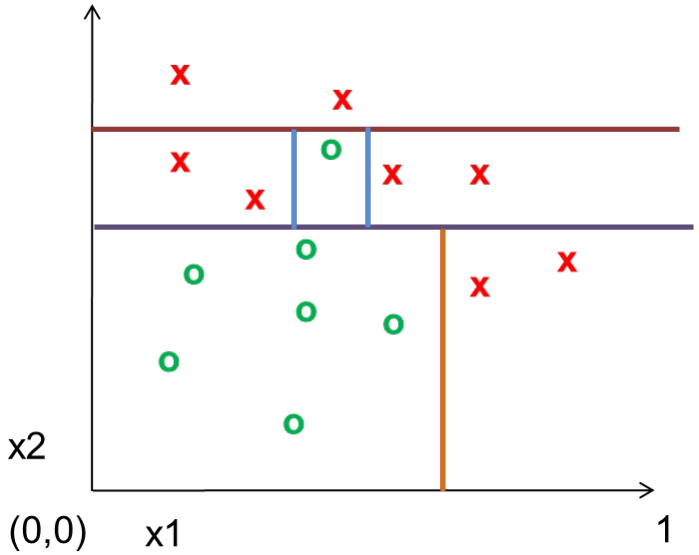






Trees (L12)

- Decision/regression trees learn to split up the feature space into partitions with similar values



- Entropy is a measure of uncertainty

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

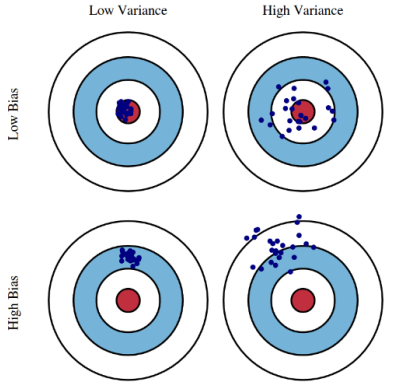
- Information gain measures how much particular knowledge reduces prediction uncertainty

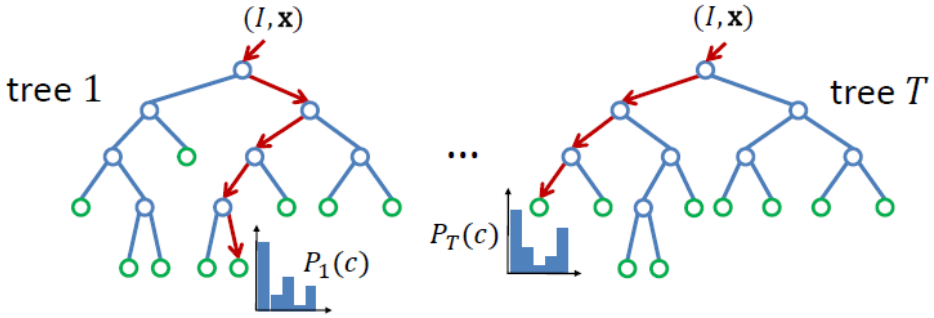
$$IG(Y|X) \quad = \quad H(Y) - H(Y|X)$$

Ensembles (L13)

- Ensembles improve accuracy and confidence estimates by reducing bias and/or variance
- Boosted trees minimize bias by fixing previous mistakes
- Random forests minimize variance by averaging over multiple different trees
- Random forests and boosted trees are powerful classifiers and useful for a wide variety of problems

$$\underbrace{E_{\mathbf{x},y,D}\left[(h_D(\mathbf{x})-y)^2\right]}_{\text{Expected Test Error}} = \underbrace{E_{\mathbf{x},D}\left[(h_D(\mathbf{x})-\bar{h}(\mathbf{x}))^2\right]}_{\text{Variance}} + \underbrace{E_{\mathbf{x},y}\left[(\bar{y}(\mathbf{x})-y)^2\right]}_{\text{Noise}} + \underbrace{E_{\mathbf{x}}\left[(\bar{h}(\mathbf{x})-\bar{y}(\mathbf{x}))^2\right]}_{\text{Bias}^2}$$





SGD (L14)

- Gradient descent iteratively takes a step in the negative gradient direction of the full objective function, to minimize a loss function
- Stochastic gradient descent (SGD) estimates the gradient using a subset of examples
  - Smaller batches require much less compute to evaluate but give a noisier estimate of the gradient
  - Faster than GD
  - Can escape local minima
- Learning rate (step size) and schedule are important factors in the speed and stability of the optimization
- Optimization problems for linear models are convex, and have a single local optimum
- MLPs and deep networks have many local optima, so are harder to optimize well