3050571 Practical Clin Data Sci

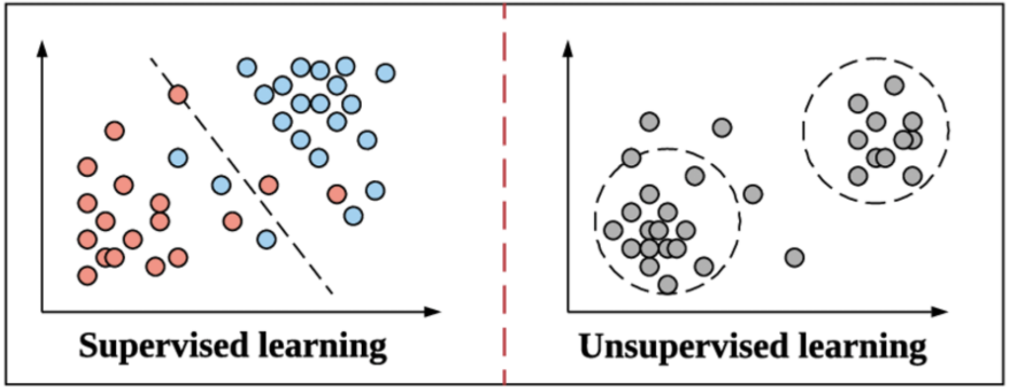
Session 10: Machine learning framework

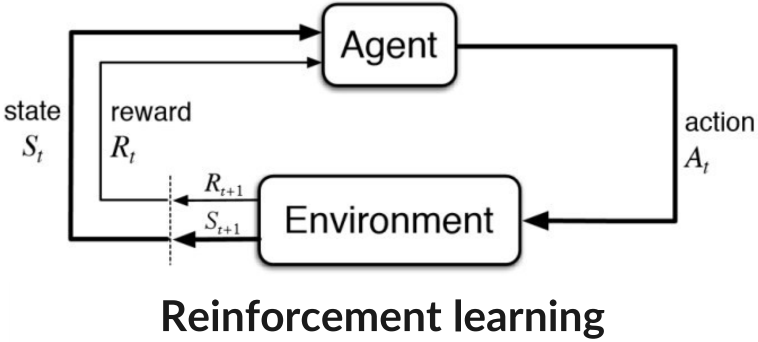
February 15, 2024

Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

Machine learning paradigms





Qian, B. et al. "Orchestrating the Development Lifecycle of Machine Learning-Based IoT Applications: A Taxonomy and Survey"

- **Supervised**: Model learns from a dataset (*x*, *y*) to predict *y* from *x*
- **Unsupervised**: Pattern recognition with no target output (only *x*)
- **Reinforcement**: Model learns by interacting and receiving feedbacks from the environment (dynamic data)

Machine learning versus human's way of thinking

Data
+
Hypotheses
Knowledge-based

→

Statistics
Likelihood, goodness-of-fit

→

Model that best explains the data

Data
+
Algorithms

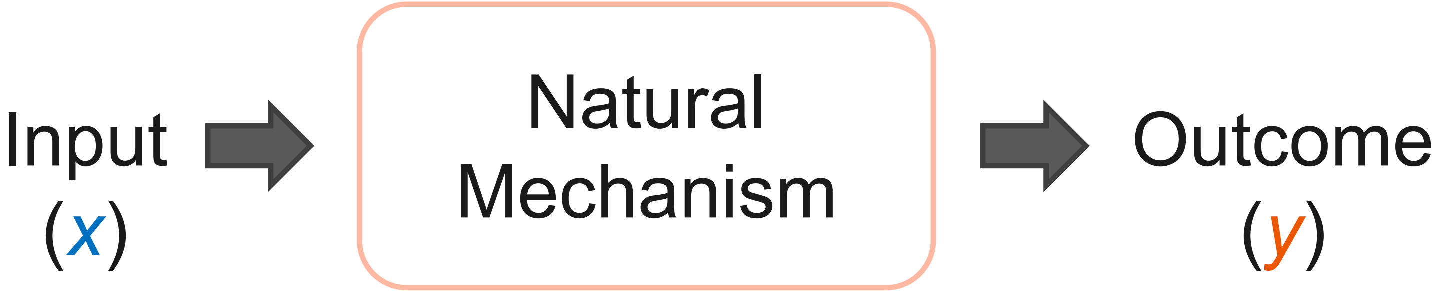
→

Machine Learning
Performance evaluation on unseen data points

→

Model that best predicts new data

Understanding versus predicting

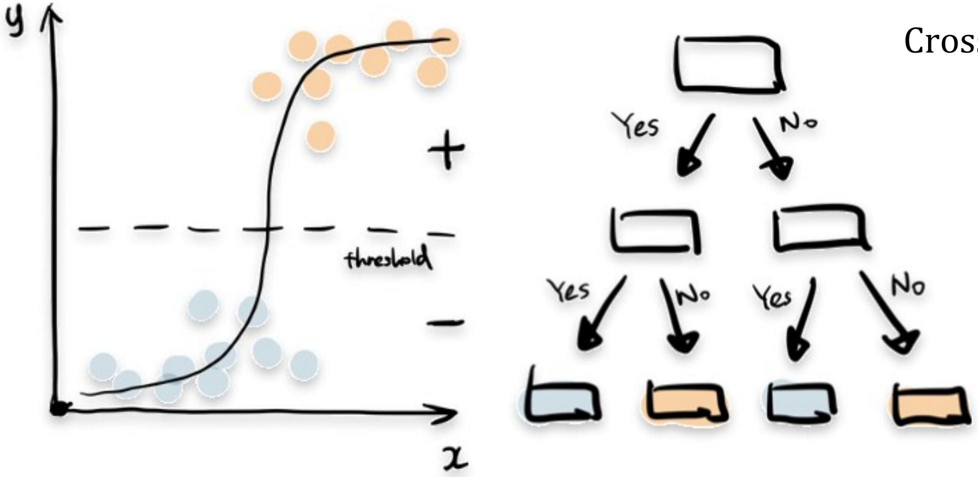


- **Statistics**: Identify the best knowledge-driven hypothesis that explain how y is generated from x
- **Supervised learning**: Find the best predictor for y from x
- **Unsupervised learning**: Use similarity among x to identify clusters and outliers, and hopefully relate them to y

Supervised learning

The cores of supervised learning

Model



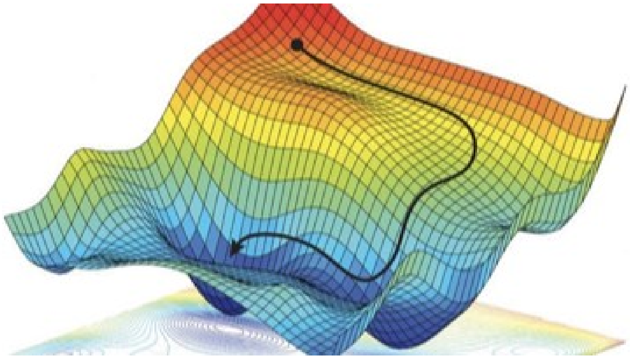
https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501

Objective / Loss Function

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad \text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \times 100$$

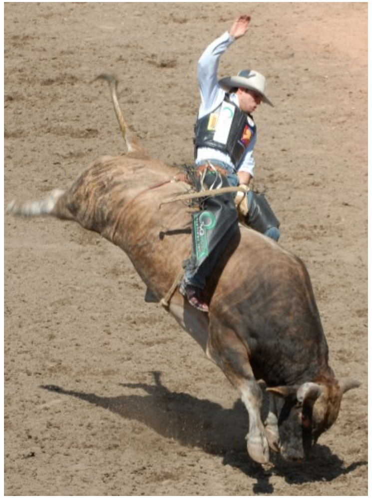
$$\text{Crossentropy} = -\frac{1}{n}\sum_{i=1}^{n} y_i \ln(\hat{y}_i) + (1 - y_i)\ln(1 - \hat{y}_i)$$

Optimization Algorithm

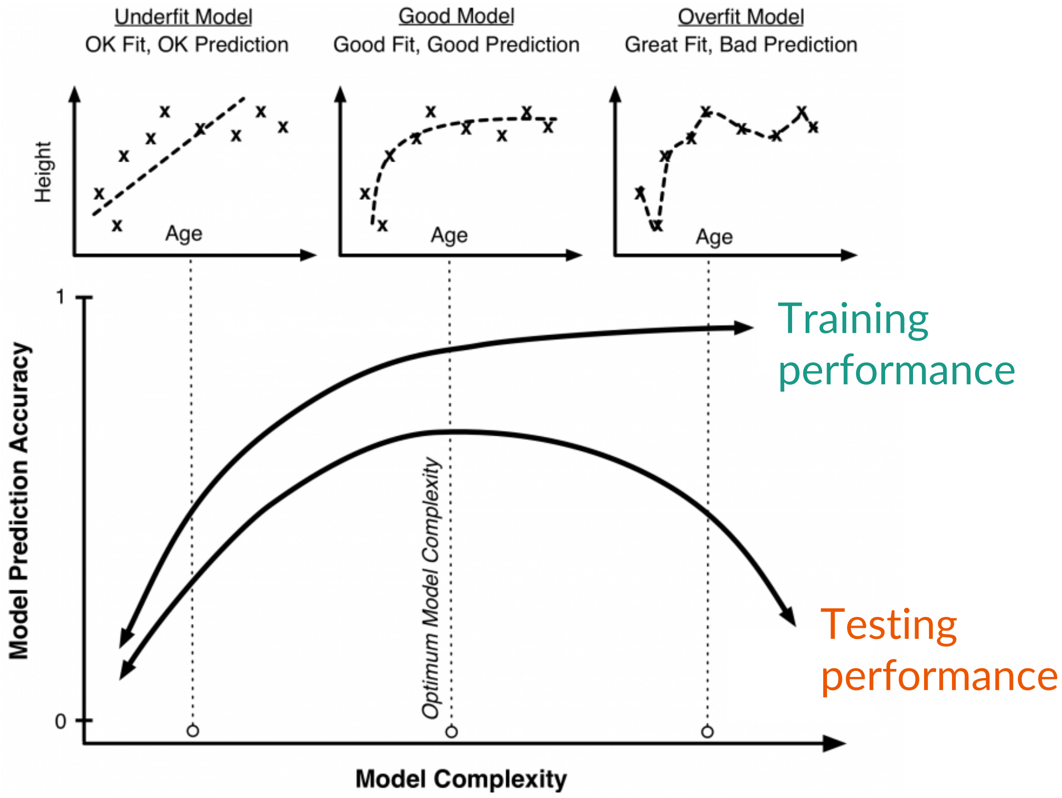


https://medium.com/analytics-vidhya/gradient-descent-b0dc1af33517

Supervised learning is all about control



https://en.wikipedia.org/wiki/Bull_riding



https://realkm.com/2018/04/23/optimization-and-complexity-the-cost-of-complexity-systems-thinking-modelling-series/

Statistical control of overfitting

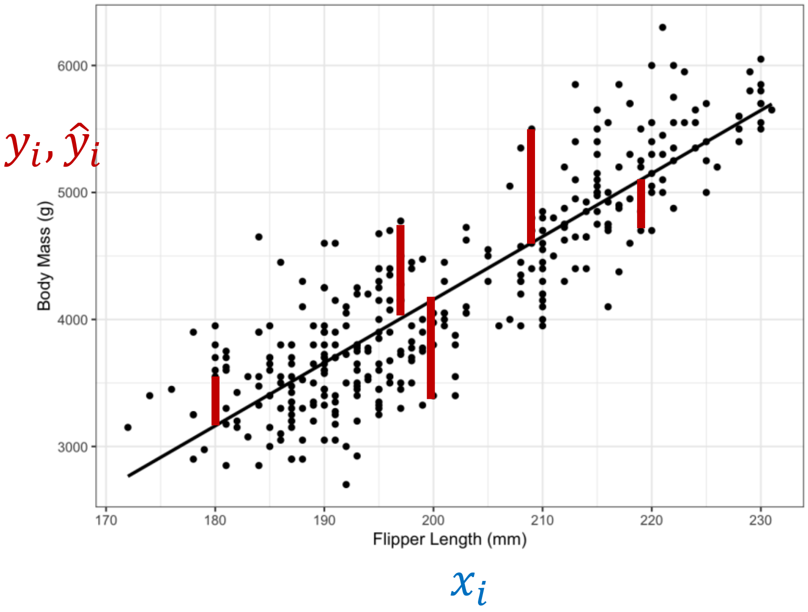
- Better model achieves higher likelihood
- Complex model has more parameters

- **Information Criterion**
  - Akaike (AIC) $= 2k - 2 \cdot \ln(\hat{L})$, where $\hat{L}$ is the likelihood
  - Bayesian (BIC) $= \ln(n)\, k - 2 \cdot \ln(\hat{L})$, where n is the sample size

- **Nested model testing**
  - Simple model has n parameters, fit the data with likelihood $\widehat{L_1}$
  - Complex model has $m > n$ parameters, fit the data with likelihood $\widehat{L_2} > \widehat{L_1}$
  - Is the improvement $\frac{\widehat{L_2}}{\widehat{L_1}}$ worth the increase in $m - n$ parameters?

Linear and logistic regression

Linear regression (Ordinary Least Square)



- Model: $\hat{y_i} = b_0 + b_1 x_i$
- Minimize MSE: $\frac{1}{n}\sum_i (y_i - [b_0 + b_1 x_i\,])^2$

- $\frac{\delta MSE}{\delta b_o} = -2\sum_i y_i - 2b_1 \sum_i x_i - 2nb_0$
- $\frac{\delta MSE}{\delta b_1} = -2\sum_i x_i y_i - 2b_1 \sum_i x_i^2 - 2b_0 \sum_i x_i$

- $b_0 = \frac{\sum xy \sum x - \sum x^2 \sum y}{(\sum x)^2 - n\sum x^2}$
- $b_1 = \frac{\sum y \sum x - n\sum xy}{(\sum x)^2 - n\sum x^2}$

Ordinary Least Square interpretation

- Observed value = True value + Normally-distributed noise
- **Assumption**: Noises are identical and independent across samples

- Model: $(y_i - \hat{y}_i) \sim N(0, \sigma^2)$
- Density: $\mathrm{P}(y_i - \hat{y}_i = \varepsilon_i \mid \sigma^2) \propto e^{\frac{-\varepsilon_i{}^2}{2\sigma^2}}$
- Likelihood: $\prod_i \mathrm{P}(y_i - \hat{y}_i = \varepsilon_i \mid \sigma^2) \propto e^{\frac{-\sum_i \varepsilon_i{}^2}{2\sigma^2}}$

- MSE: $\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_i \varepsilon_i{}^2$
- Minimizing MSE is the same as maximizing likelihood

Logistic regression

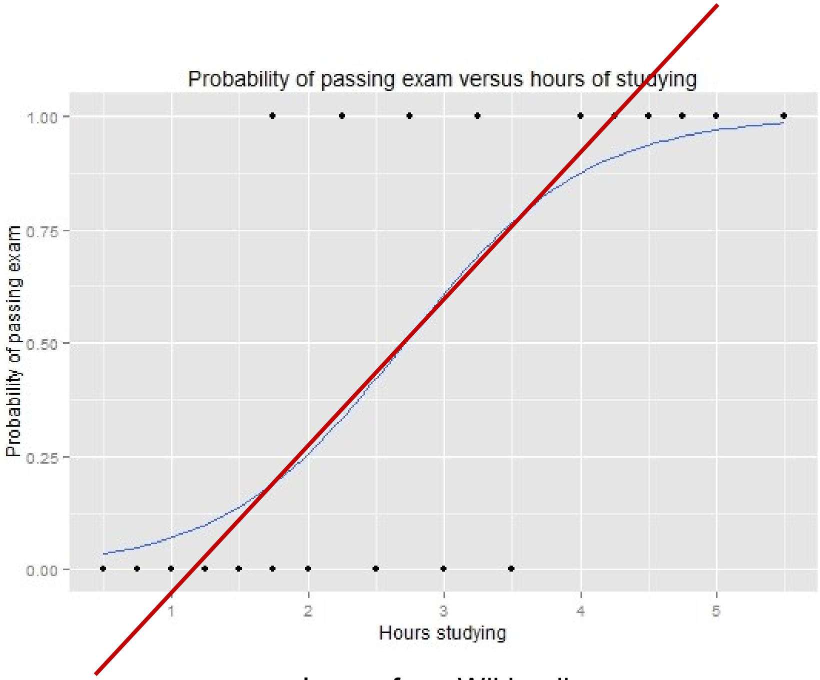


Image from Wikipedia

- Classification output = 0 or 1
- Linear regression outputs $-\infty$ to ∞

- Probability of success p
- Log odd: $\ln\left(\frac{p}{1-p}\right)$
  - $\ln\left(\frac{p}{1-p}\right) \to -\infty$ as $p \to 0$
  - $\ln\left(\frac{p}{1-p}\right) \to \infty$ as $p \to 1$

- Transform linear regression output with log odd!

Logistic regression

- Model: $\ln\left(\frac{\widehat{y_i}}{1-\widehat{y_i}}\right) = f(x_i) = b_0 + b_1 x_{i,1} + \cdots + b_n x_{i,n}$
- $\widehat{y_i} = \frac{e^{b_0+b_1x_{i,1}+\cdots+b_nx_{i,n}}}{1+e^{b_0+b_1x_{i,1}+\cdots+b_nx_{i,n}}}$
  - When $f(x_i) \to \infty$, $\widehat{y_i} \to 1$
  - When $f(x_i) \to -\infty$, $\widehat{y_i} \to 0$

- Can we keep using MSE as the loss function?
  - Brier score $= \frac{1}{N}\sum_i (y_i - \widehat{y_i})^2$
  - But this does not interpret logistic output as probability

Likelihood for logistic regression

- Likelihood: $\mathrm{P}(y_i \mid x_i) = \hat{y}_i^{\,y_i}(1-\hat{y}_i)^{1-y_i}$
  - y_i is either 0 or 1
  - When y_i is 0, the likelihood is $1-\hat{y}_i$
  - When y_i is 1, the likelihood is $\hat{y}_i$

- Log likelihood: $y_i \ln(\hat{y}_i) + (1-y_i)\ln(1-\hat{y}_i)$
  - This is the cross-entropy loss function!
  - Maximizing likelihood is the same as minimizing cross-entropy

Regularization of linear model

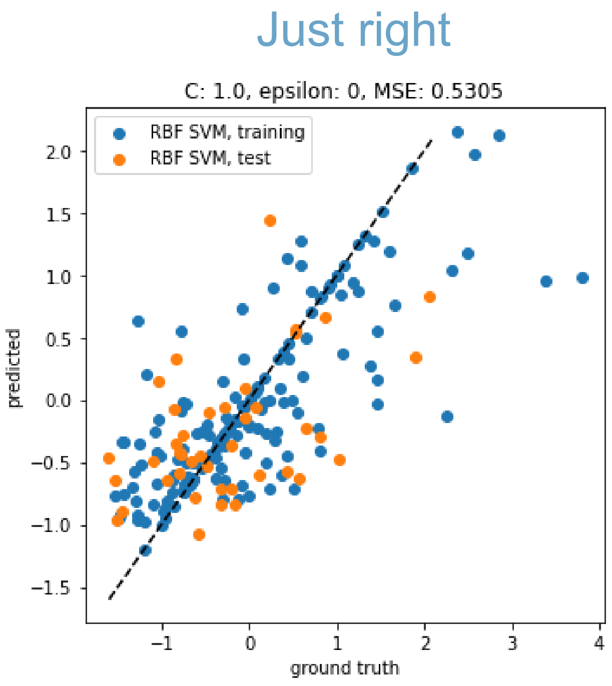
- L1 regularization (LASSO): MSE + $\alpha \sum_k |b_k|$
- L2 regularization (Ridge): MSE + $\alpha \sum_k {b_k}^2$

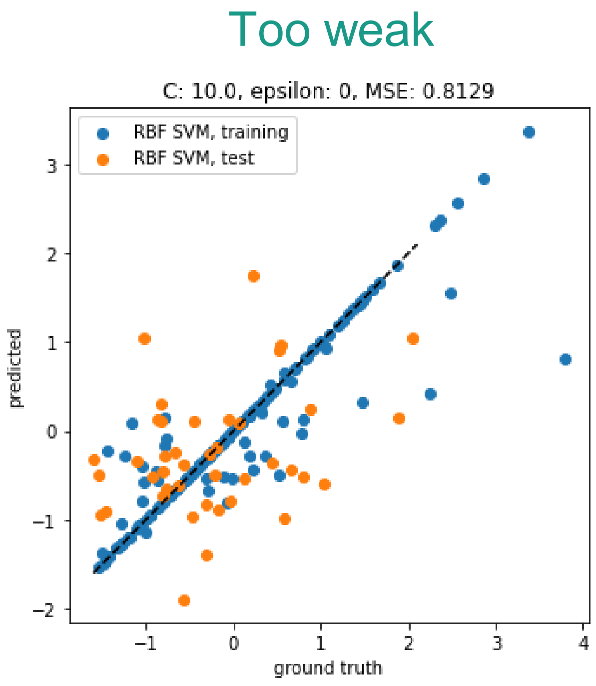
- α is the hyperparameter that controls the regularization strength

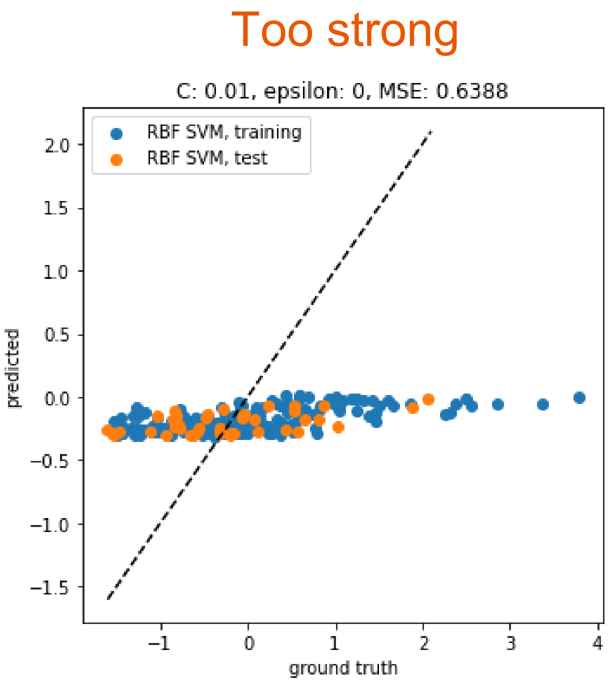
- Hyperparameter must be tuned for every dataset!

- Elastic Net = L1 + L2

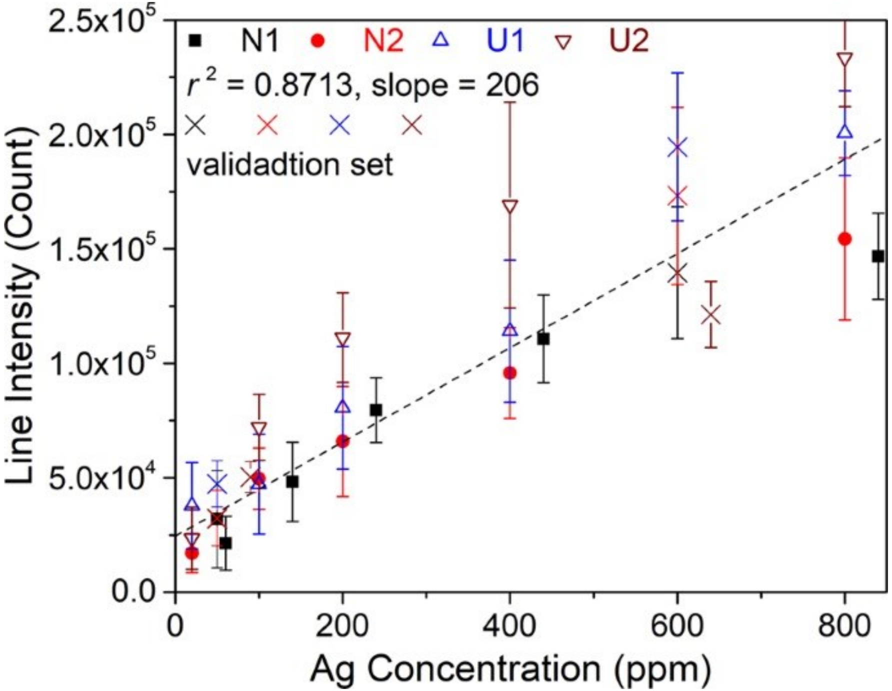
Tuning regularization strength







Weighted regression



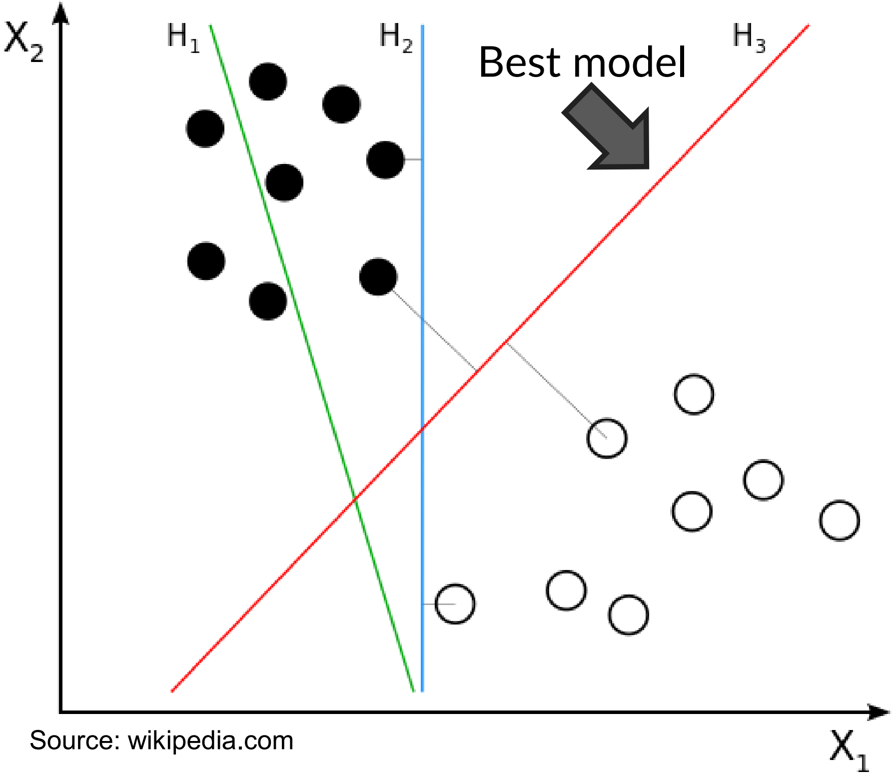
- Different data point can be weighted

- Weighted error = $\sum_i w_i (y_i - \widehat{y_i})^2$

Sun, C. et al. “Machine Learning Allows Calibration Models to Predict Trace Element Concentration in Soils with Generalized LIBS Spectra” Scientific Reports (2019)

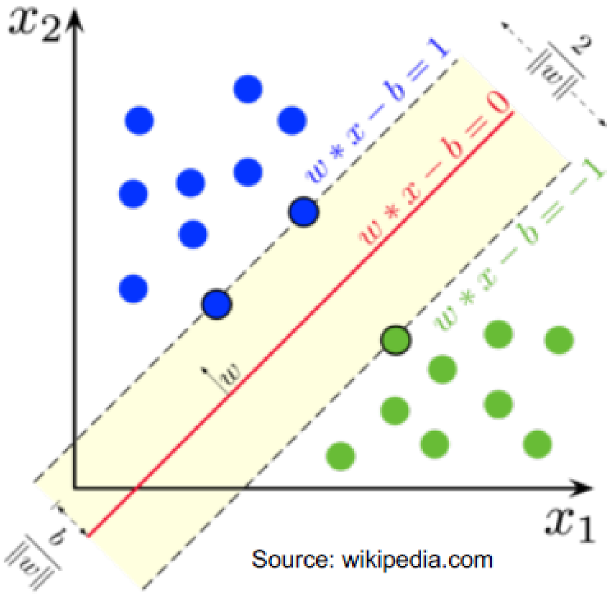
Support Vector Machine

Margin as a measure of model quality



Source: wikipedia.com

Margin as a regularization



Source: wikipedia.com

Hyperplane equation

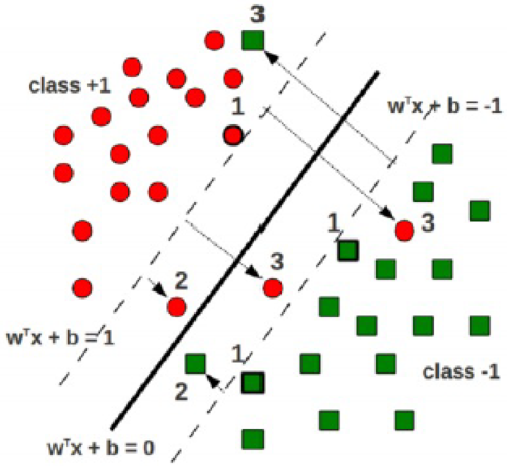
$$w_1 x_1 + \cdots + w_n x_n - b = 0$$

Scale the space so that the nearest data points on each side of the hyperplane satisfies

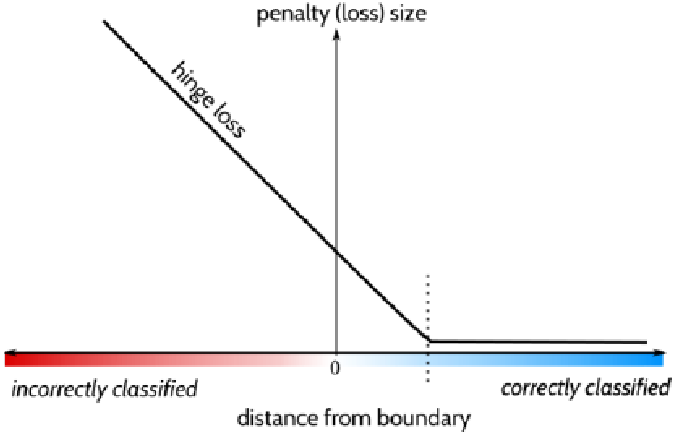
$$w_1 x_1 + \cdots + w_n x_n - b = \pm 1$$

Then, margin $= \frac{2}{\|w\|_2}$ where $\|w\|_2$ is the L-2 norm of *w*

Hinge loss



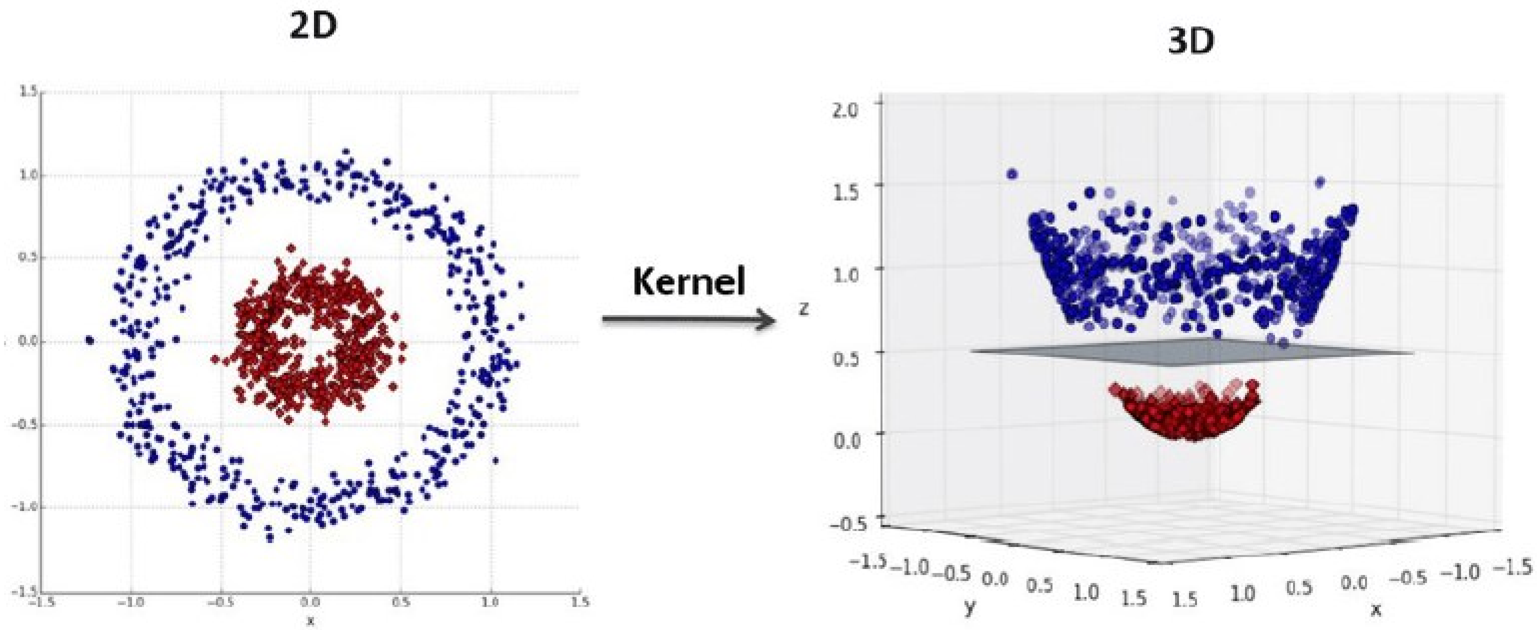
Source: cs.umd.edu/cmsc422 class



Source: towarddatasciences.com

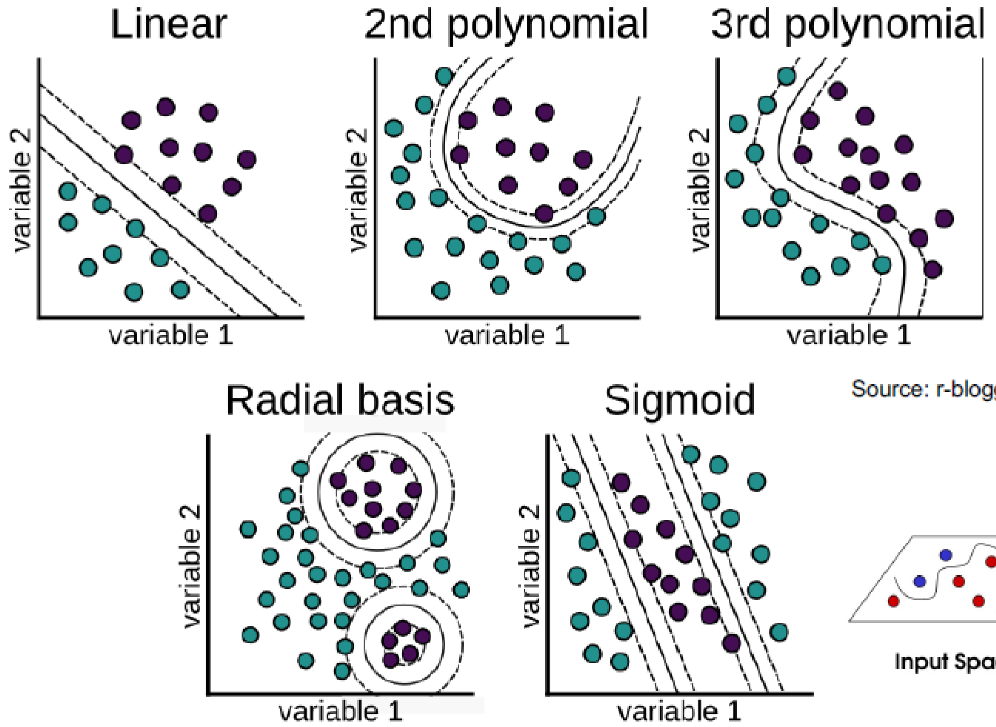
- **Hinge loss**: $\max(0, 1 - y_i(w \cdot x_i - b))$
  - Penalize misclassification and data points lying within the margin

Kernel as feature engineering

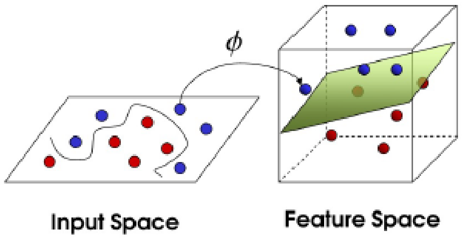


- Transform (x, y) to (x, y, $x^2 + y^2$)

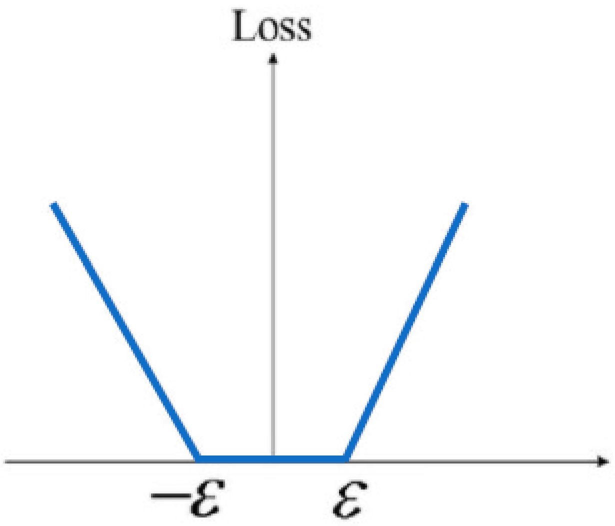
Nonlinear models from linear technique



Source: r-bloggers.com



SVM for regression = ignore small error



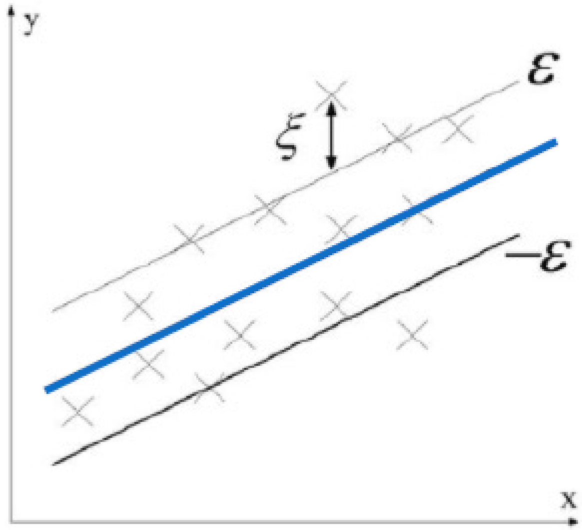
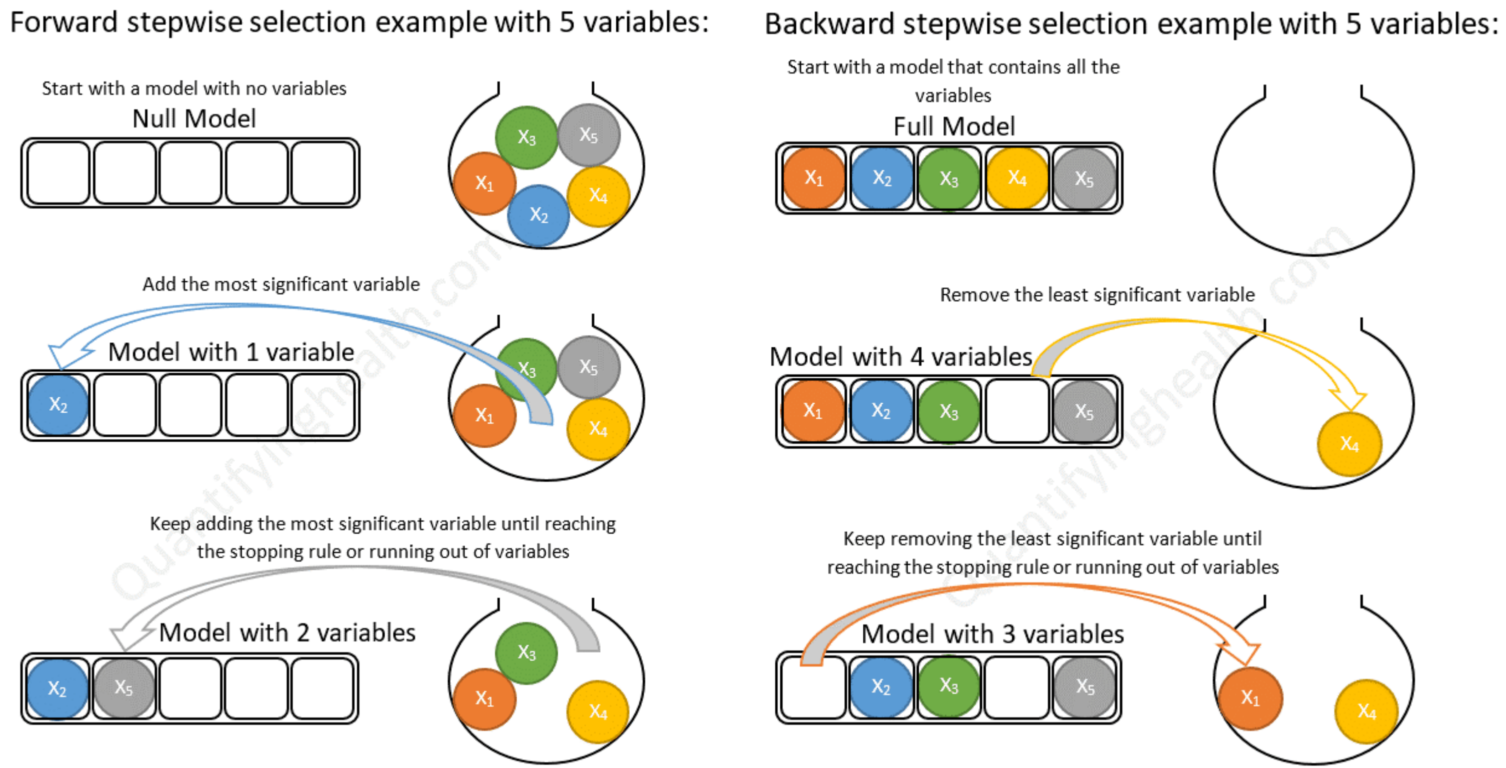


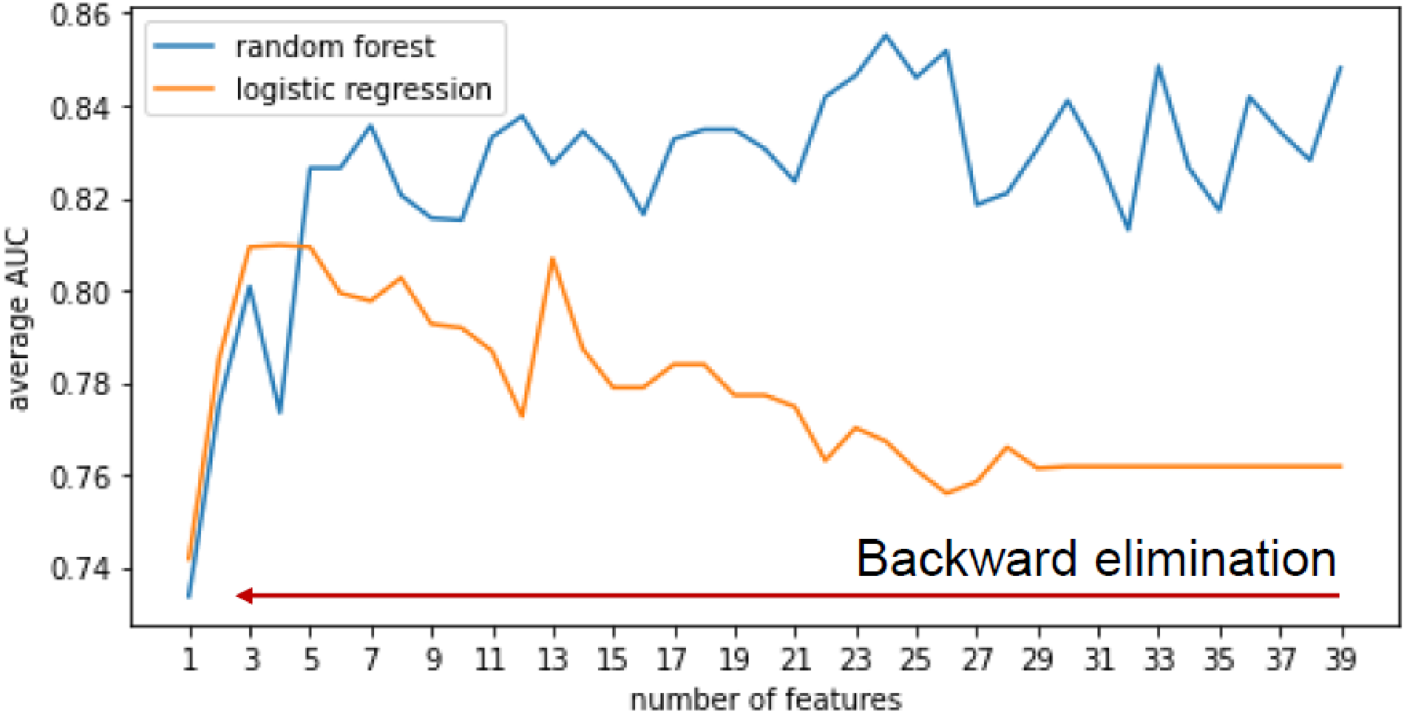
Image from https://slideplayer.com/slide/15044351/

Feature selection

Using all features can be harmful

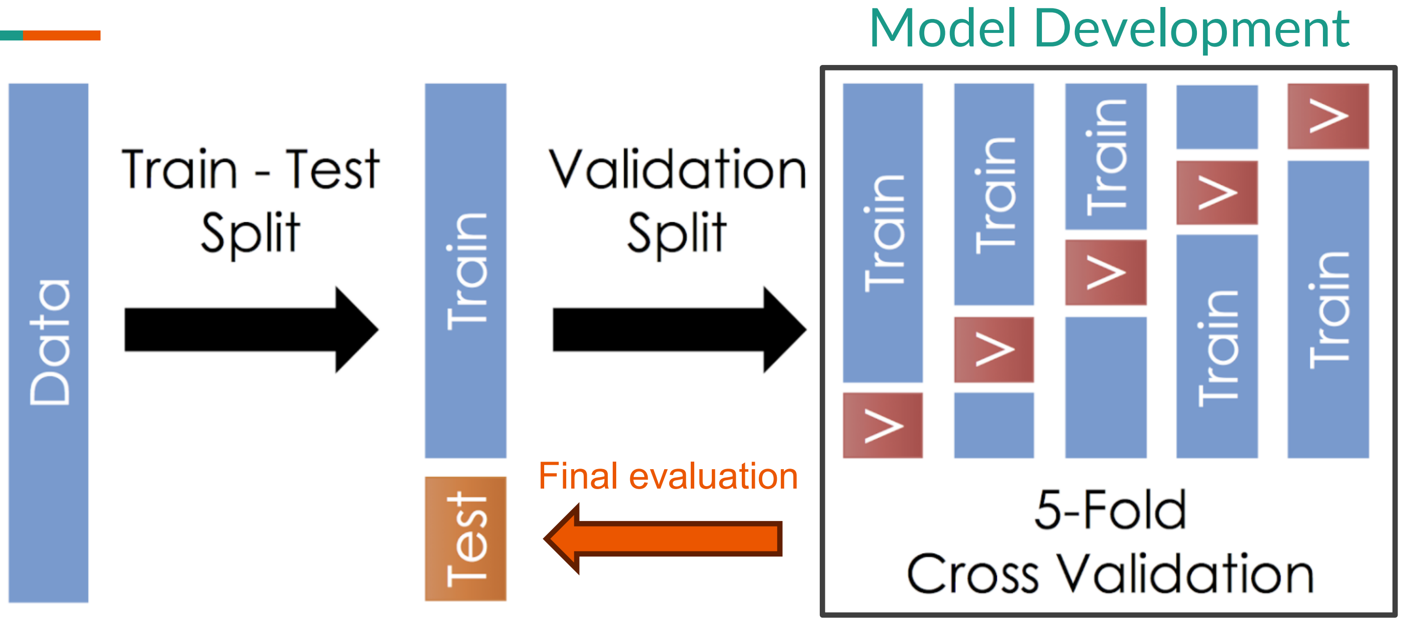


Not every model can handle large number of features



Model evaluation

Train-Val-Test

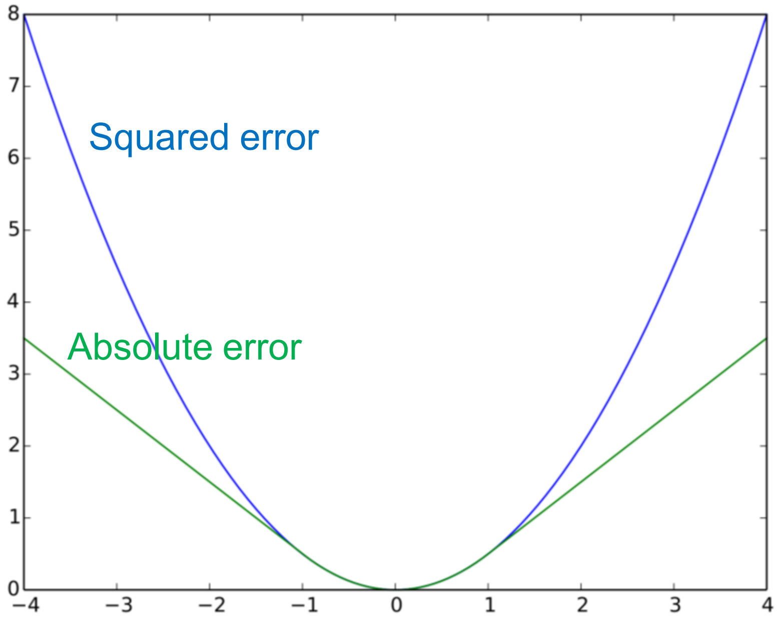


Source: medium.com

- **Training** data determines the best coefficients / weights
- **Validation** data determine the best hyperparameters
- **Test** data determine performance on new datasets

Performance metrics: regression

- Mean Square Error
- Mean Absolute Error
- Mean Absolute Percentage Error
- R^2 (Coefficient of Determination)



Performance metrics: classification

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive
		Predicted < 0.5	Predicted > 0.5

- Accuracy = (TN + TP) / total
- Precision = TP / (TP + FP) = Positive predictive value

- Recall = TP / (TP + FN) = Sensitivity
- Specificity = TN / (TN + FP)

Performance metrics can be misleading

Good

- Accuracy = (25 + 340) / 400 = 91%
- Specificity = 340 / 350 = 97%

Bad

- Precision = 25 / (25 + 10) = 71.4%
- Sensitivity = 25 / 50 = 50%

- Why is accuracy so high while sensitivity and precision are low?

	Predict YES	Predict NO
Known YES	25	25
Known NO	10	340

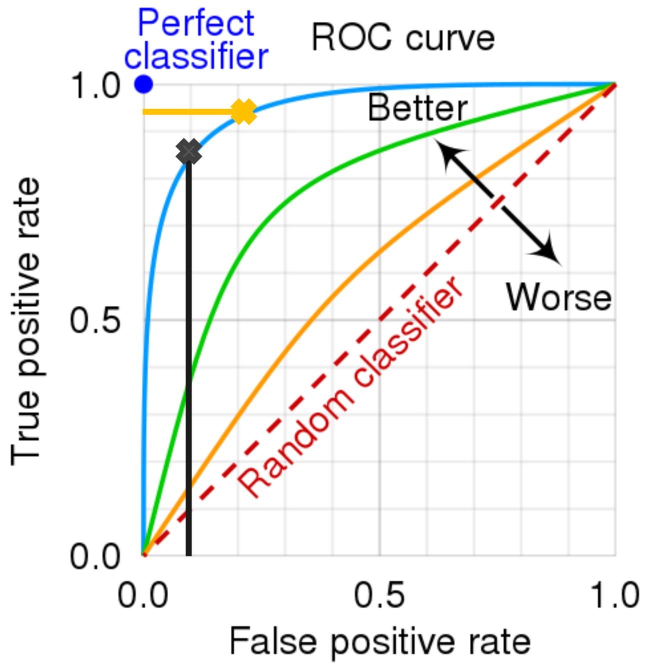
Metrics must match the question

Would you want to use this model if:

- YES = Patient will benefit from a high-risk surgery
- YES = Patient will be allergic to a given drug
- YES = Patient is at high risk of AKI

	Predict YES	Predict NO
Known YES	25	25
Known NO	10	340

Threshold-free metrics



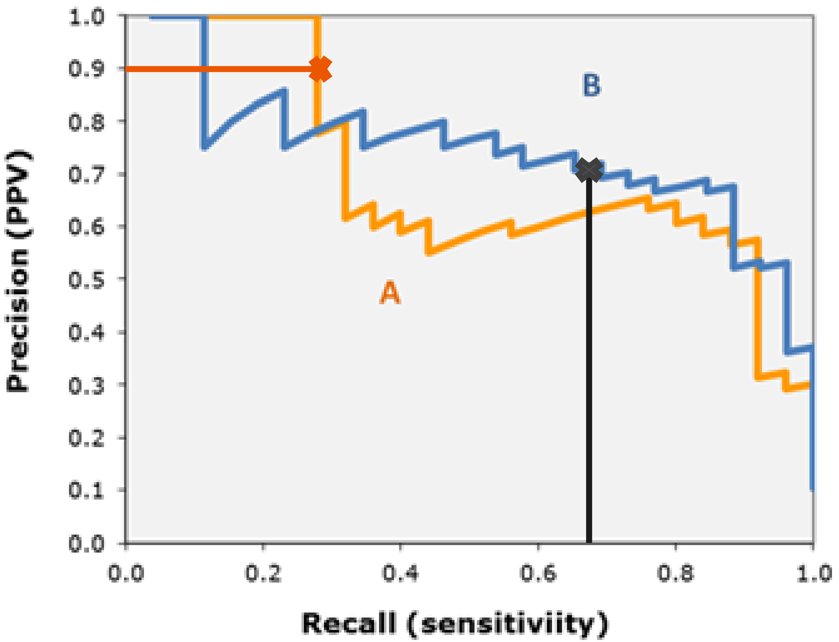
https://commons.wikimedia.org/wiki/File:Roc_curve.svg

- Sensitivity-specificity at every output threshold

- Area under the ROC curve (AUROC, AUC)
  - Random guess = 0.5
  - Perfect model = 1.0

- Pick threshold based on use case
  - Specificity >0.9
  - Sensitivity >0.9

Precision-Recall curve



https://acutecaretesting.org/en/articles/precision-recall-curves-what-are-they-and-how-are-they-used

- The best model can depend on use case

Any questions?

See you on February 22nd