Machine Learning
Lecture 2: Linear Regression

Vladislav Goncharenko
Harbor, 2021

Recap

Lecture 1: Intro to ML

- ML thesaurus
- Main problem statements (so far)
  - Supervised
    - Classification
    - Regression
  - Unsupervised
    - Dimensionality reduction
    - Clustering
- Maximum Likelihood Estimation (MLE)
- Naïve Bayes classifier
- kNN

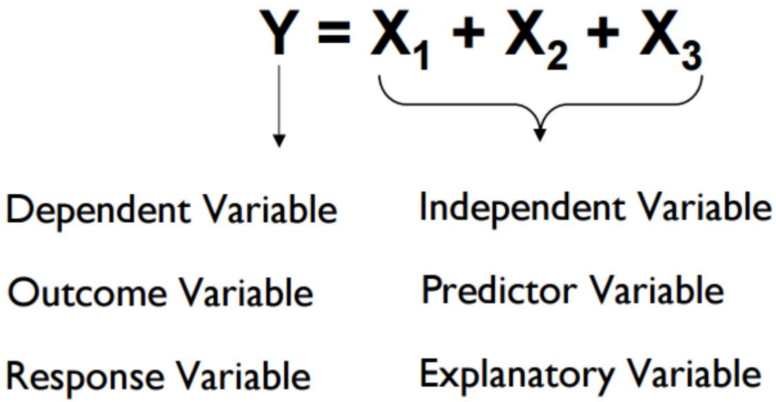
Outline

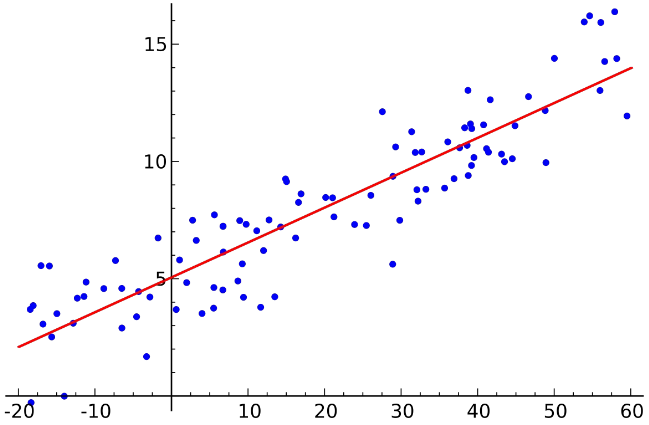
- Linear Models overview
- Regression problem statement
- Linear Regression analytical solution
    - Gauss-Markov theorem
    - Instability
- Regularization
    - L2 aka Ridge
        - Analytical solution
    - L1 aka LASSO
        - Weights decay rule
    - Elastic Net
- Metrics in regression
- Model building cycle

Linear Models overview

Linear Models overview

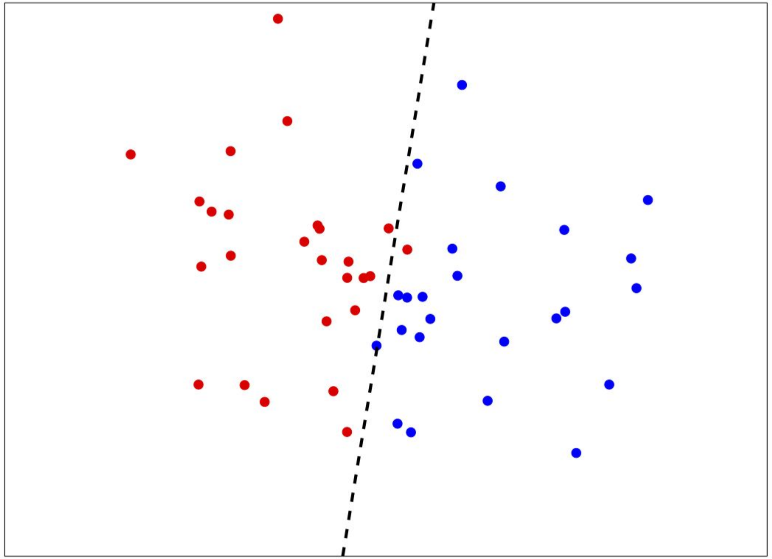
- Regression models





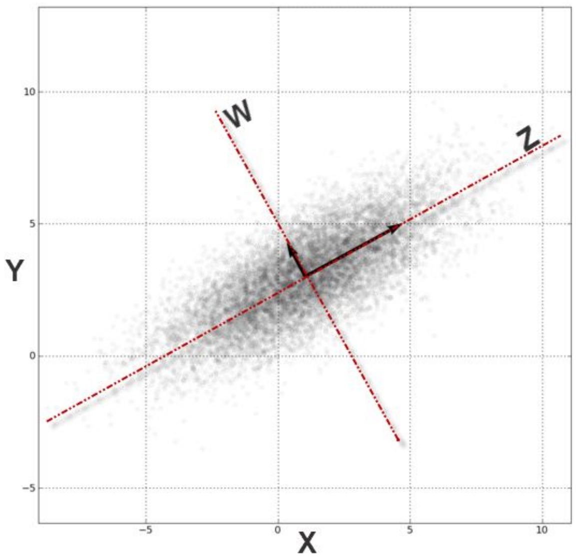
Linear Models overview

- Regression models
- Classification models



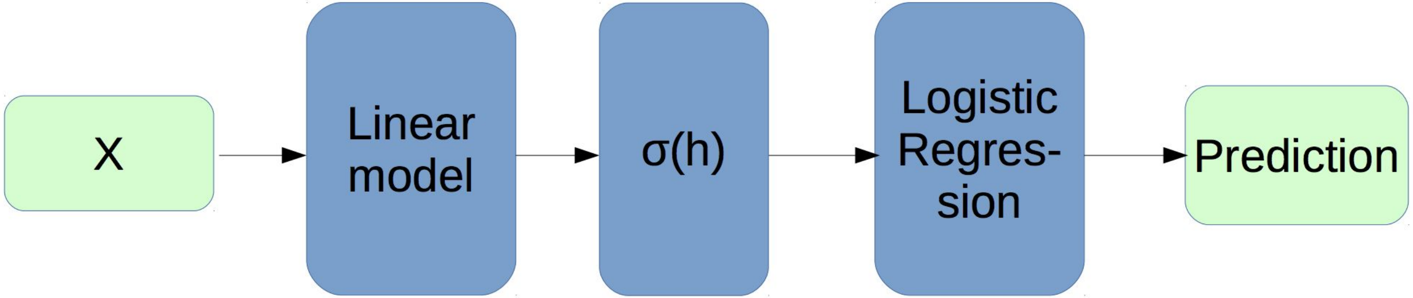
Linear Models overview

- Regression models
- Classification models
- Unsupervised models



Linear Models overview

- Regression models
- Classification models
- Unsupervised models
- Building block of other models (ensembles, NNs, etc.)



a simple neural network

Linear Regression

Regression model

Can be written in the form

$$\mathbb{E}(Y|X) = f(X)$$

or equivalently

$$Y = f(X) + \varepsilon$$

Linear Regression model

When estimator is linear

$$f_w(x) = w_0 + \sum_{i=1}^{p} w_i x_i \equiv x^T w$$

regression gets linear

Note: x and w are supposed to include bias term (conventional notation)

$$w = (w_0, w_1, \ldots, w_n)^T$$
$$x = (1, x_1, \ldots, w_n)^T$$

Linear Regression problem

Observed objects

$$(x^i, y^i), i = 1, \ldots, n$$
$$x^i \in R^p, y^i \in R$$

Matrix form of data

$$X = [x^1, \ldots, x^n]^T, X \in R^{n \times p}$$
$$Y = [y^1, \ldots, y^n]^T, Y \in R^n$$

Linear Regression

$$f_w(X) = Xw = \hat{Y} \approx Y$$

Linear Regression problem

How to choose weights?

$$\text{Empirical risk} = \sum_{\text{by objects}} \text{Loss on object} \to \min_{\text{model params}}$$

$$Q(X) = \sum_{i=1}^{n} L(y^i, f_w(x^i)) \to \min_w$$

Loss functions

$$\text{MSE:} \quad L(y_t, y_p) = (y_t - y_p)^2$$

$$\text{MAE:} \quad L(y_t, y_p) = |y_t - y_p|$$

Note: MSE minimization equivalents
Maximum Likelihood Estimation
in certain conditions (e.g. Gaussian noise)

Linear Regression analytical solution

For MSE closed form solution exists

$$Q_{\mathrm{MSE}}(X) = \sum_{i=1}^{n}(y^i - f_w(x^i))^2 = ||Y - Xw||^2 = (Y - Xw)^T(Y - Xw) \to \min_w$$

$$\begin{aligned}\nabla_w Q(X) = \nabla_w(Y^TY - (Xw)^TY - Y^TXw + (Xw)^TXw) = \\ = 0 - Y^TX - Y^TX + 2w^TX^TX = 0\end{aligned}$$

$$w^* = (X^TX)^{-1}X^TY$$

Gauss-Markov theorem

$$Y = f(X) + \varepsilon$$

$$\mathbb{E}(\varepsilon_i) = 0 \quad \forall i$$

$$\mathrm{Var}(\varepsilon_i) = \sigma^2 < \inf \quad \forall i$$

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j$$

Minimizing MSE loss gives

Best Linear Unbiased Estimation (BLUE)

(Estimator with minimal Variance from all unbiased estimators)

$$w^* = (X^T X)^{-1} X^T Y$$

$$\mathbb{E}(w^*) = w_{\text{true}}$$

$$Var(w^*) = min$$

Instability

$$w^* = \boxed{(X^T X)}^{-1} X^T Y$$

What if this matrix is singular?
e.g. strongly correlated features

Numerical inversion would be unstable

```
w_true
```

```
array([ 2.68647887, -0.52184084, -1.12776533])
```

```
w_star = np.linalg.inv(X.T.dot(X)).dot(X.T).dot(Y)
w_star
```

```
array([   2.68027723, -186.0552577 ,  184.41701118])
```

L2 Regularization

How to fix instability?

Add 100% invertible matrix

$$w = (X^T X + \lambda^2 I)^{-1} X^T Y$$

Turns out that this value is optimal solution for a penalized (by L2 norm of w) loss function

$$L_2 = ||Y - Xw||_2^2 + \lambda^2 ||w||_2^2$$

Derivation is identical to vanilla Linear Regression case discussed above

This type of regularization is called Tikhonov regularization or Ridge regression or L2 regularization

Note: regularization constraints model weights and decrease them

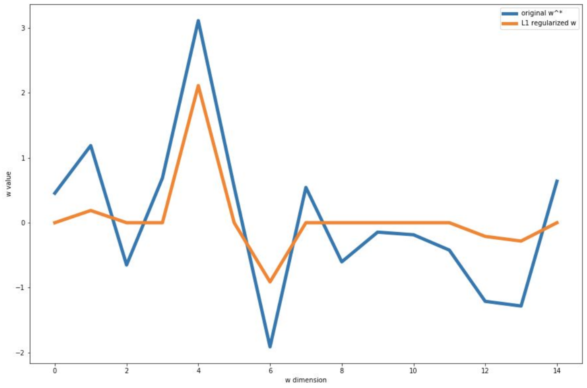
L1 Regularization

What if we add L1 norm of w to our loss? This technique is called LASSO

$$L_1 = ||Y - Xw||_2^2 + \lambda^2 ||w||_1$$

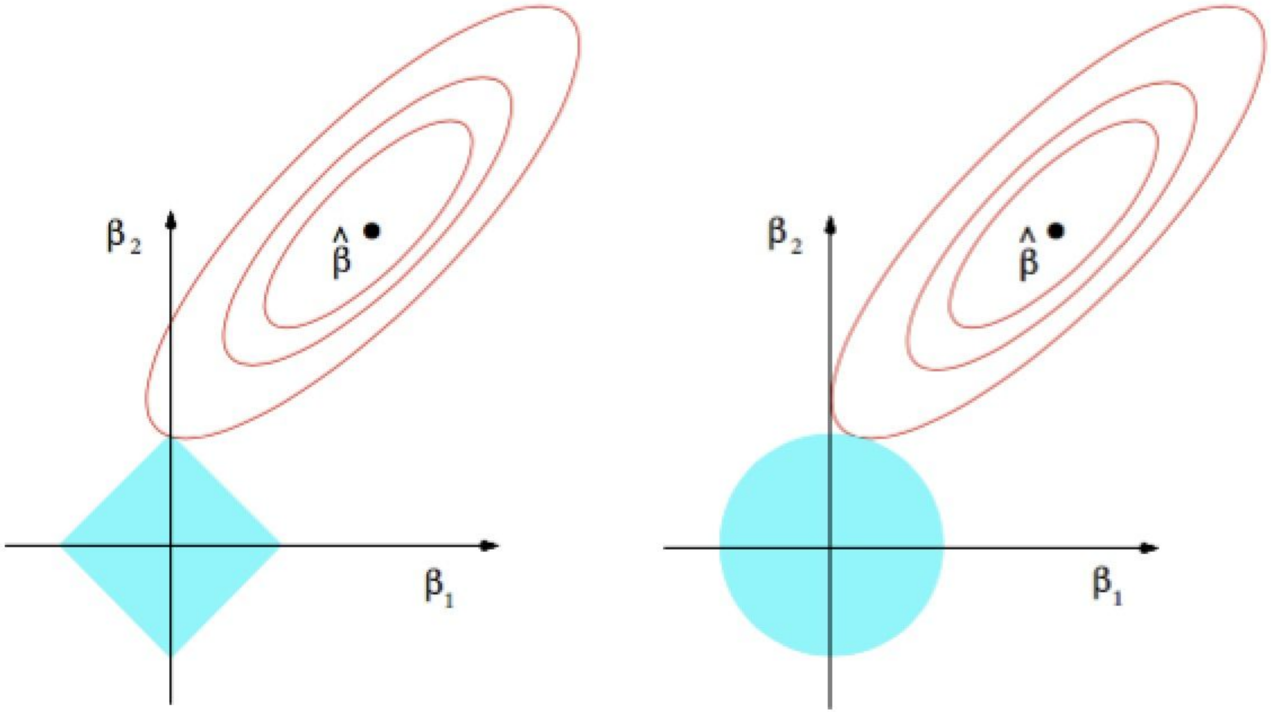
For this case there is no such an elegant solution as for L2 regularization, however solution exists for orthonormal design (see [Spokoiny's book] p.173)

$$\widehat{\theta}_j = \begin{cases} (\widetilde{\theta}_j - \lambda)_+ & \widetilde{\theta}_j \geq 0, \\ -(|\widetilde{\theta}_j| - \lambda)_+ & \widetilde{\theta}_j < 0 \end{cases}$$



Thus this type of regularization performs implicit feature selection

Regularizations geometrical interpretation



ElasticNet Regularization

Applying both types of regularization also works

$$L_{EN} = ||Y - Xw||_2^2 + \lambda_1^2||w||_1 + \lambda_2^2||w||_2^2$$

Metrics in regression

Metrics in regression

- MSE - Mean Square Error
- MAE - Mean Absolute Error
- RMSE - Root Mean Square Error
- MAPE - Mean Absolute Percentage Error
- SMAPE - Symmetric Mean Absolute Percentage Error
- R2 - “R squared” aka coefficient of determination
- … (any combination you like)

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$\mathrm{SMAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{2\cdot|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|}$$

Model building cycle

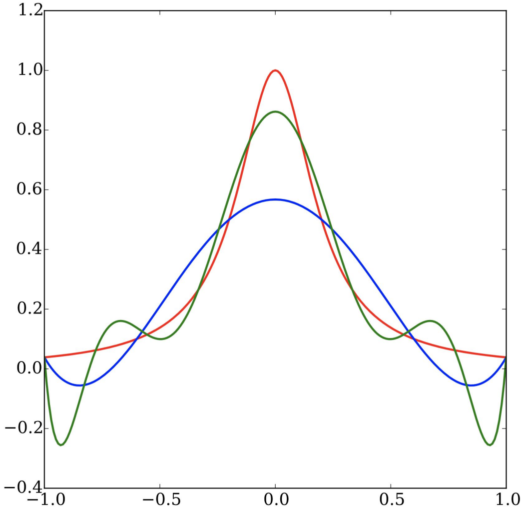
Parameters vs Hyperparameters

	Parameters	Hyperparameters
Changes	Optimized during training	Fixed before training
Choice depends on	Training set	Validation set
kNN	None	#neighbours
Linear Regression	vector w	regularization

Runge's phenomenon

Runge function interpolation on uniform grid

$$f(x) = \frac{1}{1+25x^2}, x \in (-1, 1) \quad x_i = \frac{2i}{n} - 1, i = 0, \ldots, n$$



by polynomials of n-th degree

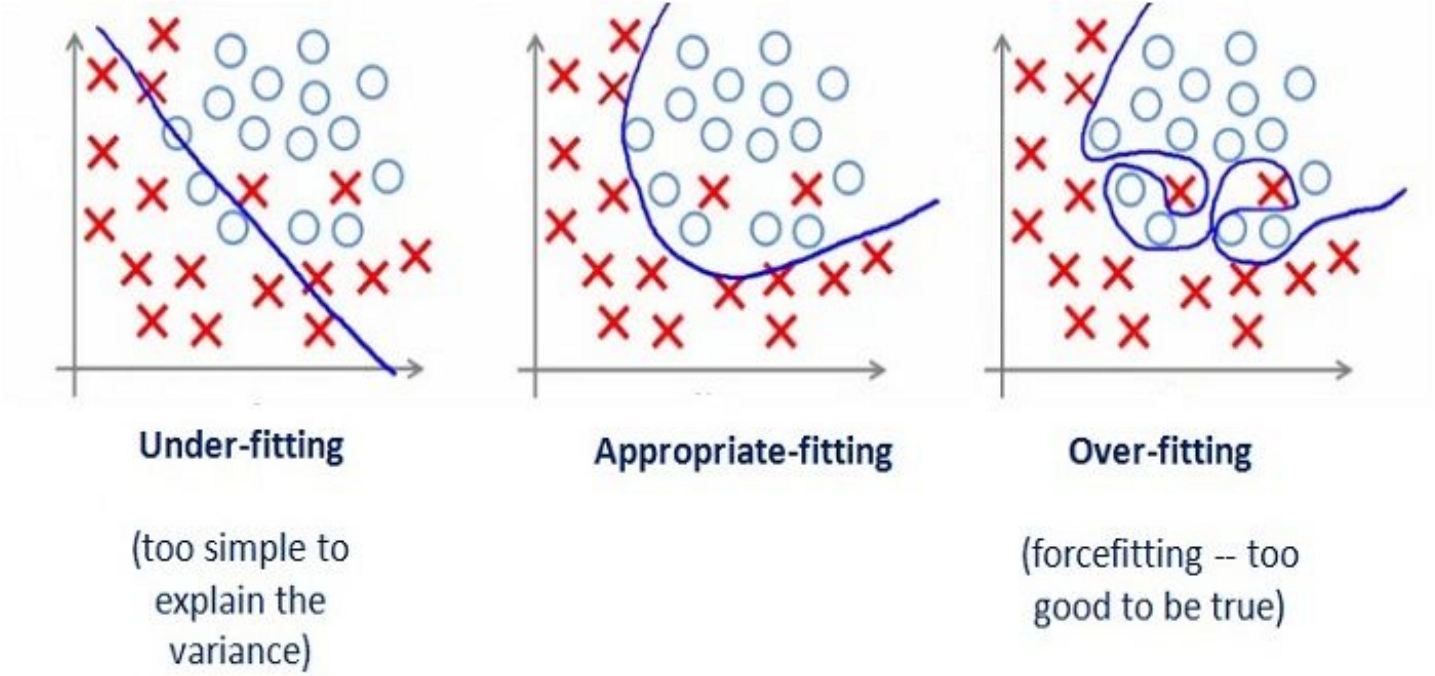
$$P_n(x) = p_n x^n + \cdots + p_1 x + p_0$$

$$P_n(x_i) = f(x_i)$$

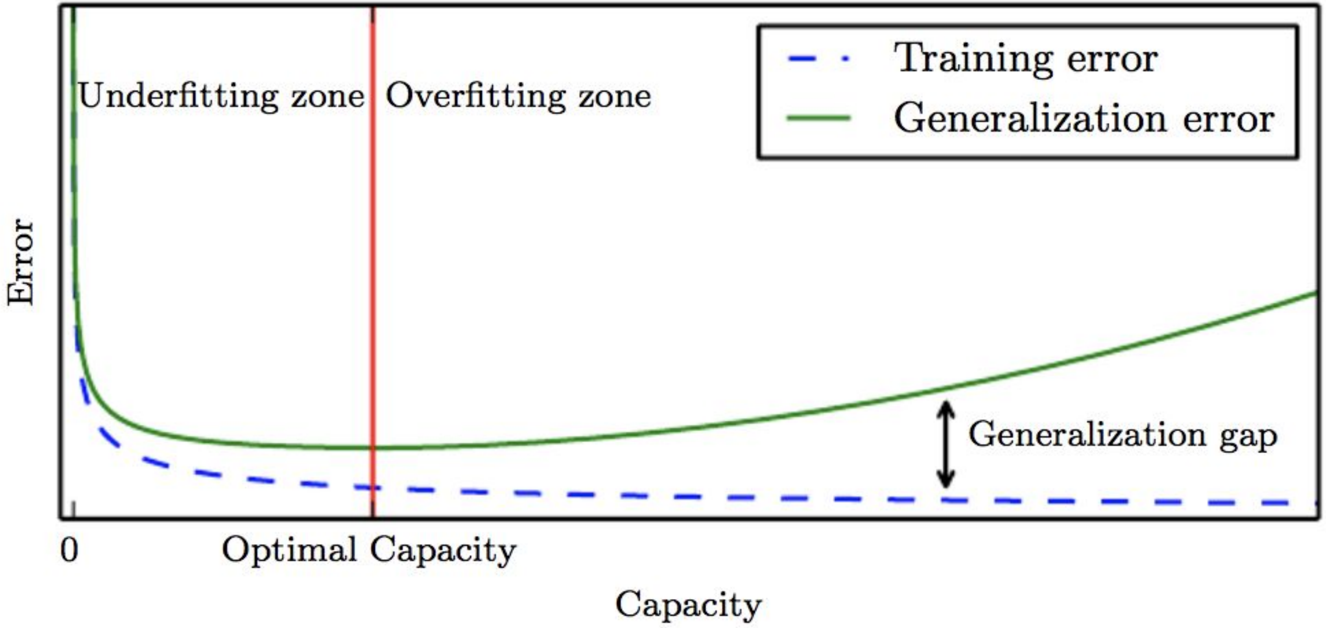
is infinitely bad on the whole interval

$$\lim_{n\to\infty}\left(\max_{-1\le x\le 1}|f(x) - P_n(x)|\right) = +\infty$$

Underfitting vs. Overfitting



Underfitting vs. Overfitting



Revise

- Linear Models overview
- Regression problem statement
- Linear Regression analytical solution
  - Gauss-Markov theorem
  - Instability
- Regularization
  - L2 aka Ridge
    - Analytical solution
  - L1 aka LASSO
    - Weights decay rule
  - Elastic Net
- Metrics in regression
- Model building cycle

Next time

- Linear classification
- Logistic regression
- Metrics in classification

Thanks for attention

Questions?