Data Culture: Intro into Machine Learning

Generalization ability. Cross-validation. Regularization

Alexey Artemov[1,2]

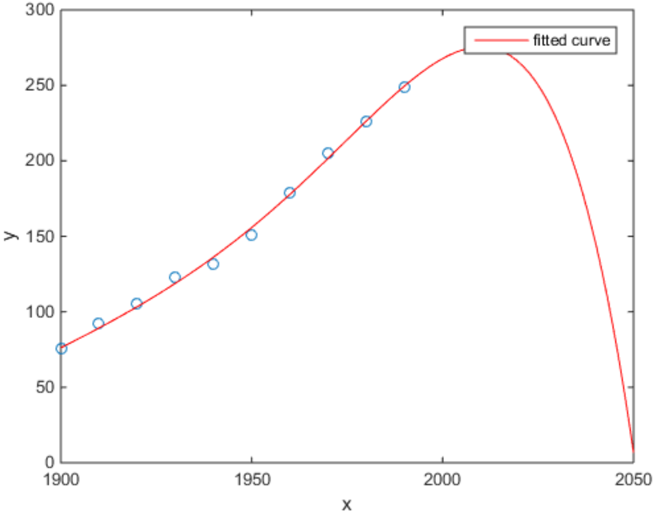
[1] Yandex LLC [2] National Research University Higher School of Economics

15 September 2017

Lecture overview

- Generalization in machine learning
- Overfitting: how to fool the linear regression
- Regularization

Motivating examples



Motivating examples

Matlab example of US census population versus time:

- A linear model is pretty good
- A quadratic model is closer
- A quartic model predicts total annihilation starting next year

(At least I sincerely hope this is an example of overfitting)

Picture credit: http://www.mathworks.com/help/curvefit/examples/polynomial-curve-fitting.html#zmw57dd0e115

Comment: The behavior of the sixth-degree polynomial fit beyond the data range makes it a poor choice for extrapolation and you can reject this fit.

Notation for today

- An unknown distribution D generates instances $(\mathbf{x}_1, \mathbf{x}_2, \ldots)$ independently
- An unknown function $f : \mathbb{X} \to \mathbb{Y}$ generates responses $(y_1, y_2, \ldots)$ for them such that $y_i = f(\mathbf{x}_i), i = 1, 2, \ldots$
- The machine learning problem: choose a plausible hypothesis $h : \mathbb{X} \to \mathbb{Y}$ from the hypothesis space $\mathbb{H}$
- The error of a hypothesis h is the deviation from the true f measured by the loss function (an example for regression):

$$Q(h, X^\ell) = \frac{1}{\ell} \sum_{i=1}^{\ell} (f(\mathbf{x}_i) - h(\mathbf{x}_i))^2$$

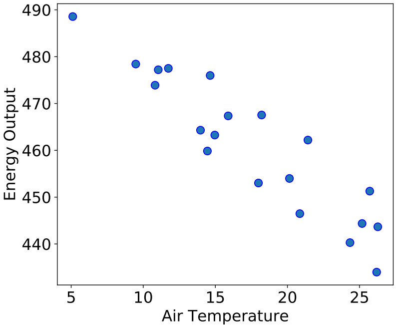
- **Learning:** the search for the optimal hypothesis $h \in \mathbb{H}$ w. r. t. the fixed loss function

Univariate linear regression

- A single feature (regressor) x: Air Temperature
- A single dependent variable y: Energy Output
- Training set $X^\ell = \{(x_i, y_i)\}_{i=1}^{20}$
- The regression model:

$$y_i = h(x_i; \mathbf{w}) + \varepsilon_i$$

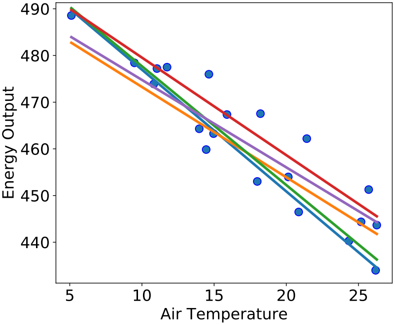
- Linear model: $y_i = w_1 x_i + w_0 + \varepsilon_i$
- The goal: given X^ℓ, find $\mathbf{w} = (w_1, w_0)$



Univariate linear regression

- Which fit to choose?
- With the linear model being fixed, depends on the data and the *loss function!*
- Mean square (L2) loss (MSE):

$$Q(h, X^\ell) = \frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - h(x_i))^2$$



Univariate linear regression

- With the loss fixed, the linear problem reduces to optimization:

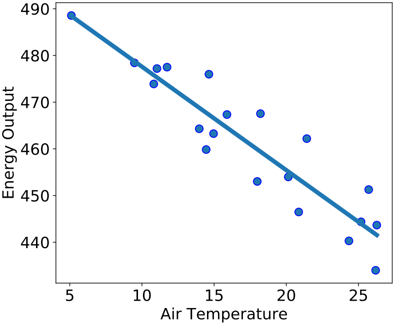
$$\frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - w_1 x_i - w_0)^2 \to \min_{(w_0, w_1)\in\mathbb{R}^2},$$

to which an analytical solution is available

$$\widehat{w}_1 = \frac{\sum_{i=1}^{\ell}(x_i - \mu_x)(y_i - \mu_y)}{\sum_{i=1}^{\ell}(x_i - \mu_x)^2},$$

$$\widehat{w}_0 = \mu_y - \widehat{w}_1 \mu_x$$

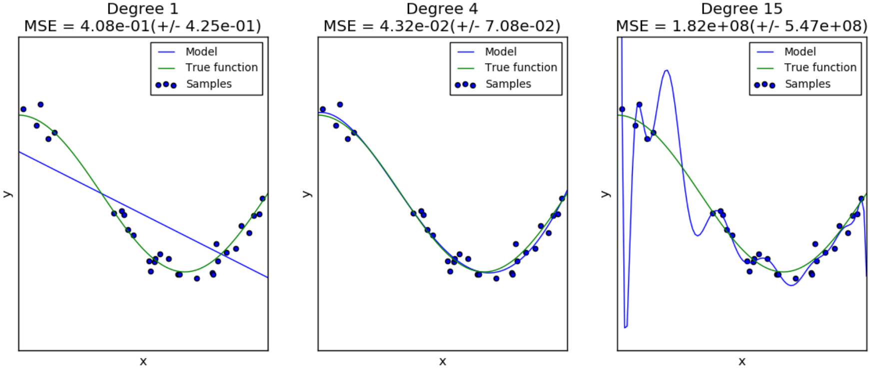
with $\mu_x = \frac{1}{\ell}\sum_{i=1}^{\ell} x_i, \quad \mu_y = \frac{1}{\ell}\sum_{i=1}^{\ell} y_i$



Generalization and overfitting

- Training set memorization: for seen $(\mathbf{x}, y) \in X^\ell$, $h(\mathbf{x}) = y$
- Generalization: equally good performance on both new and seen instances
- How to assess model's generalization ability?
- Consider an example:
    - $y = \cos(1.5\pi x) + \mathcal{N}(0, 0.01)$, $x \sim \text{Uniform}[0, 1]$
    - Features: $\{x\}$, $\{x, x^2, x^3, x^4\}$, $\{x, \dots, x^{15}\}$
- How well do the regression models perform?

Polynomial fits of different degrees



Model validation and selection

- We have free parameters in models:
  - polynomial degree d, subset of features in multivariate regression, kernel width in kernel density estimates, ...
- **Model selection:** how to select optimal hyperparameters for a given classification problem?
- **Validation:** how to estimate true model performance?
- Can we use entire dataset to fit the model?
- *Yes, but* we will likely get overly optimistic performance estimate
- **The solution:** rely on held-out data to assess model performance

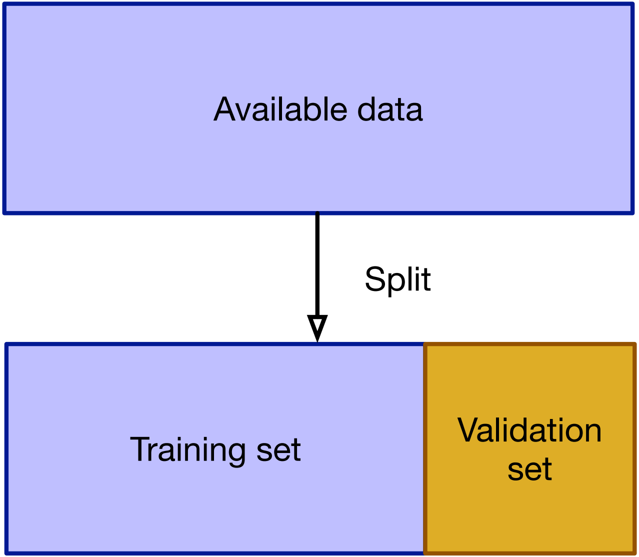
Assessing generalization ability: train/validation

- Split training set into two subsets:

$$X^\ell = X^\ell_{\mathsf{TRAIN}} \cup X^\ell_{\mathsf{VAL}}$$

- Train a model h on X^ℓ_{TRAIN}
- Evaluate model h on X^ℓ_{VAL}
- Assess quality using $Q(h, X^\ell_{\mathsf{VAL}})$

Train/validation splits



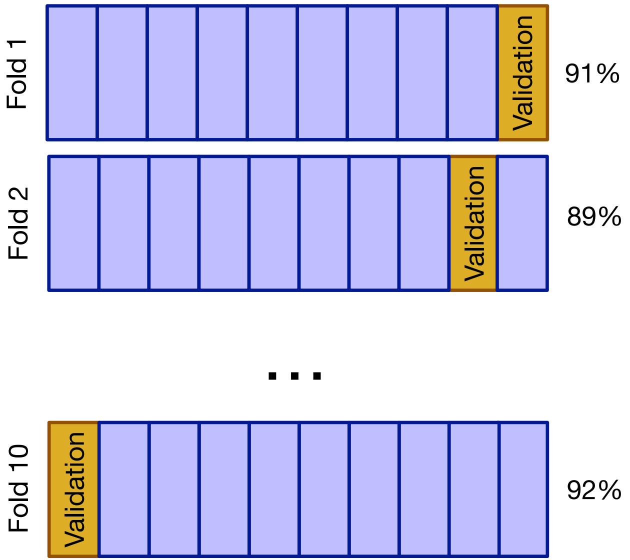
Train/validation method: drawbacks

- **Data-hungry:** we may not be able to afford the "luxury"of setting aside a portion of the dataset for testing
- **May be imprecise:** the holdout estimate of error rate will be misleading if we happen to get an "unfortunate"split

Assessing generalization ability: cross-validation

- Split training set into subsets of equal size $X^\ell = X_1^\ell \cup \ldots \cup X_K^\ell$
- Train K models $h_1, \ldots, h_K$ where each model h_k is trained on all subsets **but** X_k^ℓ
- Assess quality using $\mathrm{CV} = \frac{1}{K} \sum\limits_{k=1}^{K} Q(h_k, X_k^\ell)$ (K-fold)
- Leave-one-out cross-validation: $X_k^\ell = \{(\mathbf{x}_k, y_k)\}$

10-fold cross-validation



Cross-validation method: drawbacks

$$\mathrm{CV} = \frac{1}{K}\sum_{k=1}^{K} Q(h_k, X_k^\ell)$$

Many folds:

- **Small bias:** the estimator will be very accurate
- **Large variance:** due to small split sizes
- **Costly:** many experiments, large computational time

Few folds:

- **Cheap, computationally effective:** few experiments
- **Small variance:** average over many samples
- **Large bias:** estimated error rate conservative or smaller than the true error rate

Ad-hoc regularization: motivation

- Consider the multivariate linear regression problem with $\boldsymbol{X} \in \mathbb{R}^{d \times d}$
$$\|\boldsymbol{y} - \boldsymbol{X}\mathbf{w}\|^2 \to \min_{\mathbf{w} \in \mathbb{R}^d}$$
- Analytic solution involves computing the product $\boldsymbol{R} = (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal$
- If $\boldsymbol{X} = \text{diag}(\lambda_1, \ldots, \lambda_d)$ with $\lambda_1 > \lambda_2 > \ldots > \lambda_d \to 0$ (meaning we're in eigenbasis of $\boldsymbol{X}$) then

$$\begin{aligned}\boldsymbol{R} &= (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal = \\ &= \big(\text{diag}(\lambda_1, \ldots, \lambda_d)\text{diag}(\lambda_1, \ldots, \lambda_d)\big)^{-1}\text{diag}(\lambda_1, \ldots, \lambda_d) = \\ &= \text{diag}\Big(\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_d}\Big), \quad \text{leading to huge diagonal values in } \boldsymbol{R}\end{aligned}$$

Ad-hoc regularization: L2

- **Regularization:** replace fit with fit + penalty as in

$$Q(\mathbf{w}) \to Q_\alpha(\mathbf{w}) = Q(\mathbf{w}) + \alpha R(\mathbf{w})$$

- $R(\mathbf{w})$ is called the *regularizer*, $\alpha > 0$ – the *regularization constant*
- *Regularized* multivariate linear regression problem

$$\|\boldsymbol{y} - \boldsymbol{X}\mathbf{w}\|^2 + \alpha\|\mathbf{w}\|_2^2 \to \min_{\mathbf{w}\in\mathbb{R}^d}$$

- *Regularized* analytic solution available

$$\mathbf{w}^* = (\boldsymbol{X}^\intercal\boldsymbol{X} + \alpha\boldsymbol{I})^{-1}\boldsymbol{X}^\intercal\boldsymbol{y}$$

Why L2 regularization works

- Analytic solution: compute the regularized operator

$$\boldsymbol{R} = (\boldsymbol{X}^\intercal \boldsymbol{X} + \alpha \boldsymbol{I})^{-1} \boldsymbol{X}^\intercal$$

- If $\boldsymbol{X} = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ with $\lambda_1 > \lambda_2 > \ldots > \lambda_d \to 0$ (meaning we're in eigenbasis of $\boldsymbol{X}$) then

$$\begin{aligned}\boldsymbol{R} &= (\boldsymbol{X}^\intercal \boldsymbol{X} + \alpha \boldsymbol{I})^{-1} \boldsymbol{X}^\intercal = \\ &= \big(\mathrm{diag}(\lambda_1, \ldots, \lambda_d)\mathrm{diag}(\lambda_1, \ldots, \lambda_d) + \\ &\qquad + \mathrm{diag}(\alpha, \ldots, \alpha)\big)^{-1} \mathrm{diag}(\lambda_1, \ldots, \lambda_d) = \\ &= \mathrm{diag}\Big(\frac{\lambda_1}{\lambda_1^2 + \alpha}, \ldots, \frac{\lambda_d}{\lambda_d^2 + \alpha}\Big),\end{aligned}$$

smoothing diagonal values in $\boldsymbol{R}$

More regularizers!

- *L2 regularized* multivariate linear regression problem

$$\|\boldsymbol{y} - \boldsymbol{X}\mathbf{w}\|^2 + \alpha\|\mathbf{w}\|^2 \to \min_{\mathbf{w}\in\mathbb{R}^d}$$

- *L1 regularized* regression (LASSO)

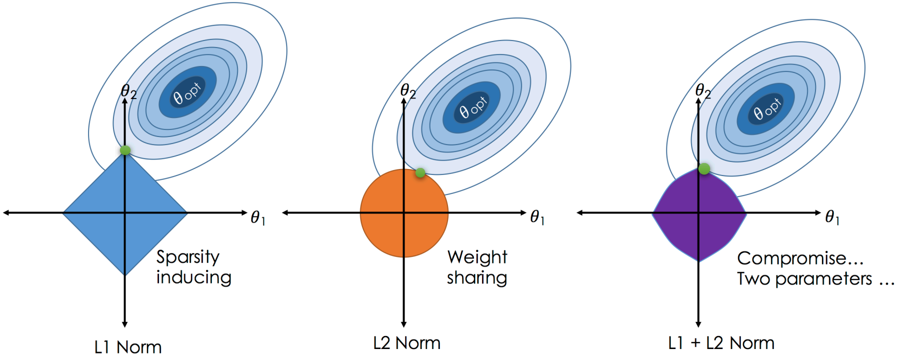
$$\|\boldsymbol{y} - \boldsymbol{X}\mathbf{w}\|^2 + \alpha\|\mathbf{w}\|_1 \to \min_{\mathbf{w}\in\mathbb{R}^d}$$

- *L1/L2 regularized* regression (Elastic Net)

$$\|\boldsymbol{y} - \boldsymbol{X}\mathbf{w}\|^2 + \alpha_1\|\mathbf{w}\|_1 + \alpha_2\|\mathbf{w}\|_2^2 \to \min_{\mathbf{w}\in\mathbb{R}^d}$$

- Convex $Q(\mathbf{w})$: unconstrained optimization $Q(\mathbf{w}) + \alpha\|\mathbf{w}\|_1$ is equivalent to constrained problem $Q(\mathbf{w})$ s.t. $\|\mathbf{w}\|_1 \leqslant C$

Geometric interpretation of regularizers



Picture credit:
http://www.ds100.org/sp17/assets/notebooks/linear_regression/Regularization.html

Another interpretation of regularizers

. 1: Large parameter space

. 2: Regularized models