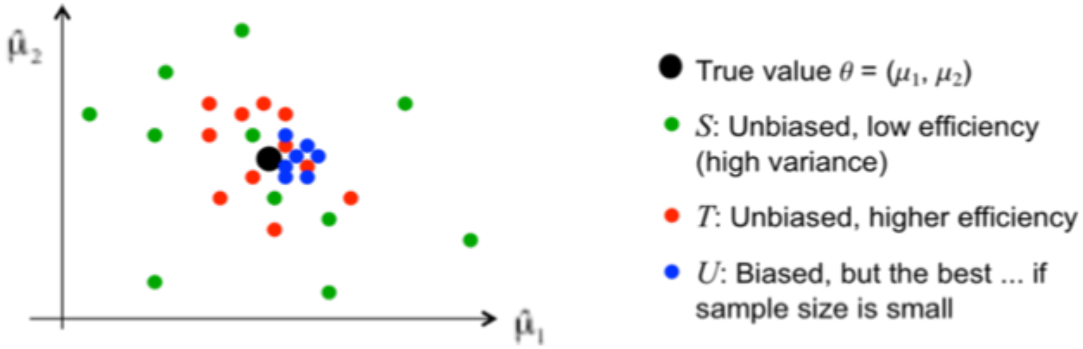
Machine Learning, Spring 2019: Exercise Sheet 7 – Solutions

Problem 1. (Visualization of the bias-variance characteristics of learning procedures). Consider a learning task where a two-parametric model of a decision function D with parameters $\theta = (\theta_1, \theta_2)$ is learnt from a training sample. For instance, θ_1 and θ_2 might be two weights for a linear decision function. Consider an repeated learning experiment where D is learnt ten times from ten different, freshly drawn training samples $(x_i^j, y_i^j)_{i = 1,\ldots,N, j = 1, \ldots, 10}$. This gives ten model estimates $\hat{\theta}[1], \ldots, \hat{\theta}[10]$. The outcome of such learning trials depends on the training algorithm that is used. Consider a scenario where three different training algorithms S, T, U are compared, leading to three times ten model estimates $\hat{\theta}^S[1], \ldots, \hat{\theta}^S[10]$; $\hat{\theta}^T[1], \ldots, \hat{\theta}^T[10]$; $\hat{\theta}^U[1], \ldots, \hat{\theta}^U[10]$. Note that each of the models $\hat{\theta}^S[1]$, etc., is a two-element parameter vector which can be conveniently plotted in a drawing plane. Furthermore let $\theta^* = (\theta^*_1, \theta^*_2)$ be the true model, that is, the parameters of the distribution from where the training samples were drawn. Assume that learning procedure S is characterized by zero bias and high variance, T is characterized by zero bias and small variance, and U has very small variance but nonzero bias. Draw a schematic plot in which you depict the 31 points θ^*, $\hat{\theta}^S[1], \ldots, \hat{\theta}^S[10]$; $\hat{\theta}^T[1], \ldots, \hat{\theta}^T[10]$; $\hat{\theta}^U[1], \ldots, \hat{\theta}^U[10]$ in different colors (black: θ^*, green: $\hat{\theta}^S[1], \ldots, \hat{\theta}^S[10]$; red: $\hat{\theta}^T[1], \ldots, \hat{\theta}^T[10]$; blue $\hat{\theta}^U[1], \ldots, \hat{\theta}^U[10]$.

Solution.



Problem 2 (Proving equation (41) from the LN) Consider a supervised learning task based on samples $(x_i, y_i)_{i = 1,\ldots,N}$ which have been obtained from random variables X and Y which take values in $\mathbb{R}^n$ and $\mathbb{R}$, respectively. Let D: $\mathbb{R}^n \to \mathbb{R}$ be a decision function. Show that the quadratic risk $E_{X,Y}[(D(X) - Y)^2]$ is minimized by

$$D_{\text{opt}}: \mathbb{R}^n \to \mathbb{R},\ D_{\text{opt}}(x) = E[Y \mid X = x].$$

Solution. We have to show that for every x,

$E[Y|X = x] = argmin_{y\in\mathbb{R}} E_{Y|X=x}[(y - Y)^2]$.

We carry out some transformations,

$$E_{Y|X=x}[(y - Y)^2] = E_{Y|X=x}[y^2 + Y^2 - 2yY] =$$

$$= E_{Y|X=x}[y^2] + E_{Y|X=x}[Y^2] - 2yE_{Y|X=x}[Y]$$

$$= y^2 + E_{Y|X=x}[Y^2] - 2yE_{Y|X=x}[Y]$$

The $argmin_{y\in\mathbb{R}}$ of this sum is independent of the middle term. We thus must find

$argmin_{y\in\mathbb{R}}\ y^2 - 2yE_{Y|X=x}[Y]$

This is a quadratic function of y which has a minimum at $y = E_{Y|X=x}[Y]$ which we also can write as $E[Y|X = x]$.

Problem 3. Consider two identically distributed, independent random variables X, Y which take values in $\mathbb{R}$. We require that X has a finite expectation, $E[X] < \infty$ (and hence, the expectation $E[Y]$ is finite too, because $E[Y] = E[X]$). Otherwise we impose no conditions on the distributions P_X, P_Y. Consider a supervised learning task with a training sample $(x_i, y_i)_{i = 1,\ldots,N}$ which has been obtained by drawing real numbers x_i and y_i with X and Y. A function D: $\mathbb{R} \to \mathbb{R}$ is trained on this sample, using linear regression (with the constant-bias-1 extension) to minimize the MSE training error. What function D will be obtained in the limit of training sample size going to infinity?

Solution. According to what we saw in Problem 1, the constant function D^*: $\mathbb{R} \to \mathbb{R}$, $D^*(x) = E[Y]$, is the function which minimizes the quadratic risk $E[(D(X) - Y)^2]$. Since D^* is an affine function, and the affine functions can be represented by the weight vectors computed by linear regression, and since linear regression finds the weight vector that minimizes the quadratic empirical risk, and since with sample size N going to infinity, the empirical risk converges to the risk, D^* will be obtained by linear regression (on asymptotically infinite-size training samples).