# SB2b HT 2017 - Problem Sheet 1 - MSc students

1. For a given loss function $L$, the risk $R$ is given by the expected loss

$$R(f) = \mathbb{E}\left[L(Y, f(X))\right],$$

where $f = f(X)$ is a function of the random predictor variable $X$.

   (a) Consider a regression problem and the squared error loss

$$L(Y, f(X)) = (Y - f(X))^2.$$

   Derive the expression of $f = f(X)$ minimizing the associated risk.

   (b) What if we use the absolute ($L_1$) loss instead?

$$L(Y, f(X)) = |Y - f(X)|.$$

2. Let $(x_i, y_i)_{i=1}^n$ be our dataset, with $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$. Linear regression can be formulated as empirical risk minimization, where the model is to predict $y$ as $x^\top \beta$, and we use the squared loss:

$$R^{\text{emp}}(\beta) = \sum_{i=1}^{n} \frac{1}{2}(y_i - x_i^\top \beta)^2$$

   (a) Show that the optimal parameter is

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

   where $\mathbf{X}$ is a $n \times p$ matrix with $i$th row given $x_i^\top$, and $\mathbf{Y}$ is a $n \times 1$ matrix with $i$th entry $y_i$.

   (b) Consider regularizing our empirical risk by incorporating a $L_2$ regularizer. That is, find $\beta$ minimizing

$$\frac{\lambda}{2}\|\beta\|_2^2 + \sum_{i=1}^{n} \frac{1}{2}(y_i - x_i^\top \beta)^2$$

   Show that the optimal parameter is given by the ridge regression estimator

$$\hat{\beta} = (\lambda I + \mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

   (c) Compare the regularized and unregularized estimators of $\hat{\beta}$ for $n = 10$ samples generated from the following model:

$$y \sim \mathcal{N}(\mathbf{X}\beta, \sigma) \tag{1}$$

   with the following parameters and a random design matrix:

```
beta = c(-.1,.3,-.5,.2,-.5,.1,.3,7)
sigma = 0.5
p = 8
n = 10

X = matrix(rnorm(n*p),n) # gives a matrix of size n x p
y = rnorm(n, mean=X%*%beta, sd=sigma)
```

use $\lambda = .1$.

Which estimator has smaller $L_2$ norm? Investigate the bias / variance tradeoff across values of $\lambda \in [0, 10]$ as follows. Generate new datasets for training and testing and make a learning curve plot comparing training and testing error like the one shown in lecture on slide 13 of Lecture 2. Put $\lambda$ on the x-axis and mean square error (prediction error) on the y-axis. Label the points on the plot corresponding to unregularized linear regression.

3. Consider two univariate normal distributions $\mathcal{N}(\mu, \sigma^2)$ with known parameters $\mu_A = 10$ and $\sigma_A = 5$ for class A and $\mu_B = 20$ and $\sigma_B = 5$ for class B. Suppose class A represents the random score $X$ of a medical test of normal patients and class B represents the score of patients with a certain disease. A priori there are 100 times more healthy patients than patients carrying the disease.

   (a) Find the optimal decision rule in terms of misclassification error (0-1 loss) for allocating a new observation $x$ to either class A or B.

   (b) Repeat (a) if the cost of a false negative (allocating a sick patient to group A) is $\theta > 1$ times that of a false positive (allocating a healthy person to group B). Describe how the rule changes as $\theta$ increases. For which value of $\theta$ are 84.1% of all patients with disease correctly classified?

4. (Exercise 2.8 in *Elements of Statistical Learning*) Compare the classification performance of logistic regression, regularized logistic regression (and optionally k-nearest neighbor classification) on the ZIP code digit image dataset, restricting to only the 2s and 3s. Investigate L1 and L2 regularization alone and in combination (the "elastic net"). Show both training and testing error for each choice. The ZIP code data are available from `https://statweb.stanford.edu/~tibs/ElemStatLearn/data.html`. You can read more about the data in Section 11.7 of *Elements of Statistical Learning*.

5. **OPTIONAL:** (Exercise 2.9 in *Elements of Statistical Learning*) Consider a linear regression model with $p$ parameters, fit with unregularized linear regression (sometimes called "least squares" or "ordinary least squares" or even just OLS) to a set of training data $(x_i, y_i)_{1 \le i \le N}$ drawn at random from a population. Let $\hat{\beta}$ be the estimator. Suppose we have some test data $(\tilde{x}_i, \tilde{y}_i)_{1 \le i \le M}$ drawn at random from the same population as the training data.

   If $R_{tr}(\beta) = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - x_i^\top \beta\right)^2$ and $R_{te}(\beta) = \frac{1}{M} \sum_{i=1}^{M} \left(\tilde{y}_i - \tilde{x}_i^\top \beta\right)^2$, prove that

   $$E[R_{tr}(\hat{\beta})] \le E[R_{te}(\hat{\beta})]$$

   where the expectation is over all that is random in each expression.