Deadline: Th. November 21th , 10:00 Drop your printed or legible handwritten submissions into the boxes at Samelsonplatz. Alternatively upload a `.pdf` file via LearnWeb. (e.g. exported Jupyter notebook)

1. Model Selection (8 points)

A. [2p] Explain how one can detect whether a model is over- or underfitting.

B. [2p] Explain how one can deal with a model that's over- or underfitting.

C. [2p] Consider a binary classification problem where each class is generated by a Normal distribution.

- 50% of the datapoints belong to class A and are distributed as $p(x \mid y = A) = \mathcal{N}(x \mid \mu_A, 1)$
- 50% of the datapoints belong to class B and are distributed as $p(x \mid y = B) = \mathcal{N}(x \mid \mu_B, 1)$

What the maximum accuracy any classifier could achieve for this problem, depending on $\delta = \mu_A - \mu_B$? (you can assume $\mu_A > \mu_B$). The minimal possible error is also known as the *irreducible error* or *Bayes error rate.*

D. [2p] Consider fitting a model on a new dataset. If we observe a very high training loss value, what does this tell us about the quality of the model? Is it over- or underfitting?

2. Bayesian Information Criterion (8 points)

The is commonly assumed in regression problems that the target variables y are generated by a deterministic function f and an additive, white noise error term ϵ, i.e.

$$y_i = f(x_i) + \epsilon_i \quad \text{where} \quad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

The goal is to recover the function f. Towards this goal, a parametric function $\hat{y}(x; \beta)$ is chosen, and the model is learned by maximizing the conditional likelihood $p(y \mid x) = \mathcal{N}(y \mid \hat{y}(x; \beta), \sigma^2)$.

A. [2p] Show that for such a model, the conditional log-likelihood has the form

$$\ell(\beta, \sigma^2) = -\tfrac{1}{2\sigma^2}\,\mathrm{RSS} - \tfrac{1}{2} N \log\bigl(2\pi\sigma^2\bigr)$$

B. [2p] Show that the maximum likelihood estimate for σ^2 is $\hat{\sigma}^2 = \mathrm{MSE}(\hat{y}) = \frac{1}{N}\|y - \hat{y}(x; \beta)\|_2^2$.

C. [1p] Conclude that, up to constant terms, for models of the kind described above the BIC is given as:

$$\mathrm{BIC}(\hat{y}) = -\tfrac{1}{2} N \log(\mathrm{MSE}(\hat{y})) - \tfrac{1}{2} D \log(N). \tag{1}$$

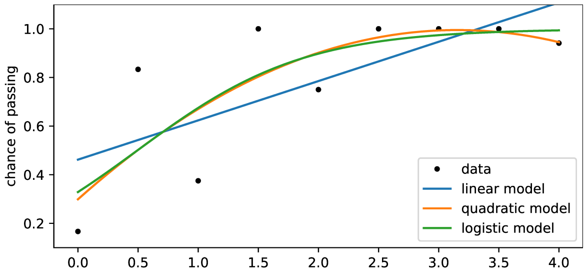
Recall that in tutorial 1, we fitted a linear regression model to predict the chance of a student passing, given the number of bonus points he obtained. We later discussed how a logistic model would have been better suited for the task. Many students also suggested a quadratic model in their submissions. Below, you find a summary of the models, their optimal parameters and a plot.

$$\hat{y}^{\text{lin}}(x) = \alpha_0 + \alpha_1 x \qquad \hat{\alpha} = \begin{pmatrix} 0.46 \\ 0.16 \end{pmatrix}$$

$$\hat{y}^{\text{quad}}(x) = \beta_0 + \beta_1 x + \beta_2 x^2 \qquad \hat{\beta} = \begin{pmatrix} 0.30 \\ 0.44 \\ -0.07 \end{pmatrix}$$

$$\hat{y}^{\text{log}}(x) = \sigma(\gamma_0 + \gamma_1 x) \qquad \hat{\gamma} = \begin{pmatrix} -0.71 \\ 1.44 \end{pmatrix}$$

(a) optimal parameters of the models



(b) plot of the models

D. [3p] Use formula (1) that we derived in parts A-C to determine which is the best model according to the BIC criterion. In contrast, which model has the lowest MSE?

3. Ridge Regression & Hyperparameter Optimization (10 points)

Many hyperparameters are discrete and thus cannot directly be trained by gradient descent. However as we will see continuous hyperparameters such as the regularization strength λ in Ridge Regression can be optimized by Gradient Descent. Assume we are given a training set (X, y) and a validation set $(\widetilde{X}, \widetilde{y})$.

$$\mathscr{L}^{\text{train}}(\beta) = \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$$
$$\mathscr{L}^{\text{val}}(\beta) = \|\widetilde{y} - \widetilde{X}\beta\|_2^2$$

And define $\hat{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}}\, \mathscr{L}^{\text{train}}(\beta)$. $\boxed{\text{Note that we restrict } \lambda \geq 0 \text{ throughout this problem.}}$

A. [2p] Show that the optimal parameters $\hat{\beta}$ of Ridge Regression satisfy the modified normal equation

$$(X^{\mathsf{T}}X + \lambda\,\mathbb{I})\hat{\beta} = X^{\mathsf{T}}y$$

B. [2p] What happens when we (erroneously) try to learn λ by updating $\lambda \leftarrow \lambda - \eta\frac{\partial}{\partial\lambda}\mathscr{L}^{\text{train}}(\beta, \lambda)$?

C. [4p] (Using jacobian layout convention). Compute the outer gradient

D. [2p] Show that if the training set is equal to the validation set, i.e. $\widetilde{X} = X$ and $\widetilde{y} = y$, then the optimal choice is no regularization at all, i.e. $\underset{\lambda}{\operatorname{argmin}}\, \mathscr{L}^{\text{val}}(\hat{\beta}(\lambda)) = 0$.