Deep Learning Worksheet 1:
Estimator Theory

Note: All questions have equal weight.

Note: Multiple answers are sometimes possible, only the correct combination of answers counts.

Submission deadline: 23:59 MEST, Friday, May 4th, 2018.

Characterization of the learning problem

We are given the observation pairs $(\mathbf{x}_t, y_t)_{t=1,\ldots,N}$ and we want to solve the following regression problem:

$$\min_{\theta} \sum_{t=1}^{N} \left(y_t - f(\mathbf{x}_t;\, \theta)\right)^2$$

Q1: Characterize the learning problem

A	B	C	D
Supervised	Semi-supervised	Unsupervised	Generative

Q2: Which of the following aspects apply to the regression problem above?

1. It minimizes the mean square error.
2. It is Tschebyscheff regression
3. It maximizes the likelihood of y_t having been emitted from $f(\mathbf{x}_t;\, \theta)$ plus a Gaussian error.

A	B	C	D	E	F	G
1 only	2 only	3 only	1 and 2	1 and 3	2 and 3	1, 2 and 3

Modeling and numerics

Suppose we observe data that has been generated by a function of the form

$$y_i = a + bx_i + cx_i^2 + dx_i^4 + \epsilon$$

where ϵ is an iid normally distributed measurement error. All coefficients are nonzero. We want to construct a linear least squares regression estimator:

$$\min_{\mathbf{w}} \sum_{t} \left(y_t - \sum_{j} w_j \phi_j(\mathbf{x}_t) \right)^2$$

Q3: Which of the following feature spaces can achieve a zero bias estimator?

A	B	C	D
$\boldsymbol{\phi} = (a, b, c, d)$	$\boldsymbol{\phi} = (x, x^2, x^4)$	$\boldsymbol{\phi} = (1, x, x^2, x^3, x^4)$	None of these

We use linear least squares (LLS) to fit a given dataset with the feature space $\boldsymbol{\phi} = (x, x^2, \cos(x), \sin(x - \frac{\pi}{2}))$ (cos and sin are taken in radians).

Q4: What can be said about the regression result?

A	B	C	D
The optimal LLS solution will have the form $\mathbf{w} = (a, b, 0, 0)$ where a and b depend on the data.	The optimal LLS solution will have the form $\mathbf{w} = (a, b, c, -c)$ where a, b and c depend on the data.	The feature correlation matrix $\mathbf{X}^\top\mathbf{X}$ is invertible.	The optimal $L2$-regularized result can be found with Ridge regression with nonzero λ parameter.

Hyperparameter selection

We fit a given dataset using Ridge regression with a polynomial model of the general form $f(x) = w_1 + w_2x + w_3x^2 + ...w_nx^{n-1}$. We want to determine the maximum polynomial order with hyperparameter optimization.

Q5: Count the number of parameters and hyperparameters

A	B	C	D	E
1 parameter, n hyperparameters	2 parameters, n hyperparameters	n parameters, 1 hyperparameters	n parameters, 2 hyperparameters	n parameters, n hyperparameters

Suppose the observation data $(\mathbf{x}_t, y_t)_{t=1,...,N}$ has not been generated from a function that can be represented by a finite order polynomial, but we still want to approximate the function with a polynomial model. After 1000 datapoints, we conduct hyperparameter optimization and conclude that polynomial order n is optimal. Now consider we instead observe 100,000 datapoints coming from the same distribution.

Q6: Which polynomial order will now likely be optimal?

A	B	C
$< n$	n	$> n$

We have training data $\mathbf{X}^{train}$ and validation data $\mathbf{X}^{val}$. We suppose that all data points from both datased have been sampled independently and identically distributed from our data source. Suppose we train a model with parameters θ and hyperparameters λ. We have an error function E that serves as a loss function.

Q7: Now we want to construct an optimal estimator that minimizes the out-of-sample error E_{out}. What is the best strategy?

A	B	C
find an unbiased estimator, and then minimize the variance within unbiased estimators.	minimize the bias in the training error and minimize the variance in the validation error	determine parameters by minimizing the loss function on the training set, determine hyperparameters by minimizing the loss function on the validation set.

We have recorded a dataset $\mathbf{X}$. We are confident that we have enough data to have a representative sample, but we are unsure if two subsequently generated datapoints are statistically independent of another. Suppose we want to train a model with fixed hyperparameter settings and get a least biased estimate of the error on data not used for the training.

Q8: Which of the following strategies is best for this purpose?

A	B	C	D
Use the first 80% of the data for training and the last 20% to compute the test error.	Shuffle the data, i.e. randomly reorder the time points, and perform A.	Perform five-fold cross-validation and use the mean validation error as test error.	Repeat B 100 times and use the mean validation error as test error.

We are given training data $(\mathbf{x}_t, y_t)_{t=1,\ldots,N}$ for fitting. We have $n = 10$ features and $N = 1000$ samples. We consider following methods, for fitting these data.

	Parameters	Training error	Validation error
Ridge regression, $\lambda = 1$	10	2.51	3.52
Kernel Ridge regression, $\lambda = 1$	1000	0.53	2.20
Two-layer Neural network with 10 hidden neurons	121	1.56	1.99
Ten-layer Deep Neural network with 10 neurons in each hidden layer	1121	0.22	2.52

Q9: Which of these models is preferable and why? Which model would you explore further and how?

A	B	C	D
Ridge regression	Kernel Ridge regression	Two-layer neural network	Ten-layer neural network