Exam for CS-E3210 - Machine Learning: Basic Principles (27.10.2016)

General Information

- Put your name and student id on EVERY page you use.
- To pass the course you must also pass the term project. Results of this examination are valid for one year after the examination date.
- Allowed equipment: calculator, pen, pencil and eraser.
- The number of points achieved for a question with M answer choices, when selecting T correct answers and F wrong answers is given by $\max\{0, 10(T-F)/M\}$.

Question 1. Consider a classification problem with three classes $\mathcal{Y} = \{C_1, C_2, C_3\}$. We observe a new point $\mathbf{x} \in \mathcal{X}$ whose likelihood under the three classes is $p(\mathbf{x}|r = C_1) = 0.01$, $p(\mathbf{x}|r = C_2) = 0.40$ and $p(\mathbf{x}|r = C_3) = 0.60$, respectively. Assume we have prior information that $p(r = C_1) = 0.98$ and $p(r = C_3) = 0.01$. According to Bayesian decision theory we should classify $\mathbf{x}$ as

A. $r = C_3$ since it was most likely to be generated from that class. ○
B. $r = C_1$ since it yields highest posterior $\frac{p(\mathbf{x}|r=C_1)p(r=C_1)}{p(\mathbf{x})} \approx 1/2$. ○
C. $r = C_1$ since it yields highest product $p(\mathbf{x}|r = C_1)p(r = C_1) \approx 1/100$. ○
D. 'reject' since no class C gives $p(r = C|\mathbf{x}) > 3/4$. ○
E. None of the other answers is correct. ○

Question 2. Consider a regression problem with scalar input $x \in \mathbb{R}$ and real-valued output $r \in \mathbb{R}$. Using polynomial basis functions $\phi_i(x) = x^i$ for linear regression $g(x) = \sum_{i=0}^{M} w_i \phi_i(x)$

A. involves $M + 1$ non-negative parameters $w_i \geq 0$, $i = 0, \ldots, M$. ○
B. gives a linear model $g(x) = w_0 + w_1 x$ if $M = 1$. ○
C. results in a non-linear predictor if $M \geq 2$. ○
D. in general tends to overfit with high M. ○
E. none of the above is correct. ○

Question 3. A popular method for choosing model complexity is cross-validation (CV), where the dataset $\mathcal{D}$ is split into training, validation and test datasets. Following statements are true:

A. We should always use 30% of data as test data. ○
B. Training and validation sets should be disjoint. ○
C. The error on the test set is an unbiased estimate for the generalisation error. ○
D. The model complexity should be selected using the error on the validation set. ○
E. In general, Leave-One-Out CV tends to overfit. ○
F. After selecting the model complexity, we can use training and validation sets to learn the model parameters. ○

Question 4. Bias-variance decomposition (BVD) refers to the identity $\mathbb{E}_{\mathcal{D}}[(g_{\mathcal{D}}(x) - f(x))^2] = \mathbb{E}_{\mathcal{D}}[(g_{\mathcal{D}}(x) - \mathbb{E}_{\mathcal{D}} g_{\mathcal{D}}(x))^2] + (\mathbb{E}_{\mathcal{D}} g_{\mathcal{D}}(x) - f(x))^2$ where $g_{\mathcal{D}}(x)$ is a regressor estimated from a dataset $\mathcal{D}$, and $f(x)$ is the assumed true function. Following statements are true:

A. BVD shows how stable a model is with respect to noise in the dataset. ○
B. BVD shows how stable a model is with respect to the model complexity. ○
C. More complex models typically yield high bias. ○
D. More complex models typically yield high variance. ○
E. A model which yields low bias tends to overfit the data. ○
F. A model yielding low bias typically incurs high variance, and vice versa. ○
G. Adding a zero-mean prior or regulariser increases bias. ○
H. There always exist models yielding zero bias and zero variance. ○
I. More complex models tend to have higher bias. ○
J. Decreasing bias is always preferred to decreasing variance. ○
K. None of the other answers is correct. ○

Question 5. In Bayesian learning, we learn a posterior distribution $p(\theta|\mathcal{D})$ of the parameters θ given the data $\mathcal{D}$. The predictive distribution $p(r|x, \mathcal{D})$ describes the distribution of an estimated response r for a new input x given the data $\mathcal{D}$. The ideal Bayesian predictive distribution uses

A. the most likely parameter values $\theta_{\text{ML}} = \operatorname{argmax}_\theta p(\mathcal{D}|\theta)$ as $p(r|x, \theta_{\text{ML}})$. ○
B. the highest posterior value $\theta_{\text{MAP}} = \operatorname{argmax}_\theta p(\theta|\mathcal{D})$ as $p(r|x, \theta_{\text{MAP}})$. ○
C. the expected likelihood value $\theta_{\text{Bayes}} = \int \theta p(\mathcal{D}|\theta) d\theta$ as $p(r|x, \theta_{\text{Bayes}})$. ○
D. the expected posterior value $\theta_{\text{Bayes}} = \int \theta p(\theta|\mathcal{D}) d\theta$ as $p(r|x, \theta_{\text{Bayes}})$. ○
E. expectation over the posterior, i.e., $p(r|x, \mathcal{D}) = \int p(r|x, \theta) p(\theta|\mathcal{D}) d\theta$. ○
F. expectation over the likelihood, i.e., $p(r|x, \mathcal{D}) = \int p(r|x, \theta) p(\mathcal{D}|\theta) d\theta$. ○
G. None of the other answers is correct. ○

Question 6. Let $\mathcal{I}$(“*statement*″) denote the indicator function which is equal to one if “statement” is true and equal to zero else. We call a classifier $h(\cdot) : \mathbb{R}^2 \to \{0, 1\}$ linear if

A. its decision boundary is a line. ○
B. its decision boundary is a circle. ○
C. it can be written as $h(\mathbf{x}) = \mathcal{I}(\mathbf{w}^T\mathbf{x} \geq w_0)$ for some $\mathbf{w} \in \mathbb{R}^2, w_0 \in \mathbb{R}$. ○
D. it can be written as $h(\mathbf{x}) = \mathcal{I}(x_1^2 + x_2^4 \geq 0)$. ○
E. None of the other answers is correct. ○

Question 7. The technique called “Bagging”

A. combines the predictions obtained for different (but related) datasets. ○
B. puts more emphasis on training examples which are predicted incorrectly. ○
C. only amounts to resampling the dataset. ○
D. None of the other answers is correct. ○

Question 8. We observe labeled data $\mathcal{D} = \{\mathbf{x}^t, r^t\}_{t=1}^N$ which we stack into the vector $\mathbf{r} = (r^1, \ldots, r^N) \in \mathbb{R}^N$ and matrix $\mathbf{X} = (\mathbf{x}^1, \ldots, \mathbf{x}^N)^T \in \mathbb{R}^{N \times d}$, respectively. The optimal linear predictor $g(\mathbf{x})$ of the real-valued output/label r^t from the multivariate input $\mathbf{x}^t \in \mathbb{R}^d$ under squared error loss $E(g(\cdot)|\mathcal{D}) := (1/N) \sum_{t=1}^N (r^t - g(\mathbf{x}^t))^2$

A. is always given by $g(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ with weight vector $\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{r}$. ○
B. is always unique, i.e., there is one and only one optimal predictor $g(\mathbf{x})$. ○
C. can be found by gradient descent for the cost function $E(g(\cdot)|\mathcal{D})$. ○
D. does not exist for some datasets. ○
E. None of the other answers is correct. ○

Question 9. The "Bootstrap" method

A. is a parametric machine learning method. ○
B. is a classification method. ○
C. allows to assess the reliability of a hypothesis. ○
D. can only be used for unlabeled data. ○
E. can be used for "Bagging". ○
F. None of the other answers is correct.