ML 1 Midterm Solutions (2015)

October 8, 2015

Questions

1. In Amsterdam, scooters are allowed to ride on the bike paths if they have a blue license plate, but are not allowed on the roads. The reverse is true for scooters with yellow license plates. The city estimates that at any given time, anywhere in the city, that a scooter on a bike path is yellow 5% of the time (i.e. the vast majority of the scooters stick to where they belong).

   One evening there is a hit-and-run accident between a scooter and a cyclist on a bike path. A witness tells police that the scooter had a yellow license plate. The police want to assess the reliability of the witness by testing him with different scooters under the same conditions the evening of the accident. The witness correctly identifies the colour of a license plate 8/10 times. In other words, if the police test the witness with a blue bike, the witness will claim they saw blue 8 of 10 tests with blue; the same is true for testing and claiming a yellow bike.

   We introduce a discrete random variable C for license plate colour that can take values y or b (yellow or blue). We are interested in the probability of the colour of a scooter's license plate *on the bike path*. We also introduce a discrete random variable W for the color that a witness claims to see that can take on values y and b (yellow or blue).

   Given this information, answer the following questions:

   (a) What is $P(C = b)$ and $P(C = y)$ on a bike path? **/1**

   Answer:

   $$P(C = b) = \frac{95}{100} \qquad P(C = y) = \frac{5}{100}$$

   (b) What is the probability that the accident was caused by a blue licensed scooter, if the witness claims it was blue? I.e. what is $P(C = b|W = b)$? **/3**

   Answer:

$$\begin{aligned} P(C = b|W = y) &= \frac{P(C = b)P(W = y|C = b)}{P(C = b)P(W = y|C = b) + P(C = y)P(W = y|C = y)} \\ &= \frac{\frac{95}{100} \cdot \frac{8}{10}}{\frac{95}{100} \cdot \frac{8}{10} + \frac{5}{100} \cdot \frac{2}{10}} \\ &= \frac{95 \cdot 8}{95 \cdot 8 + 5 \cdot 2} = \frac{760}{770} \end{aligned}$$

(c) If there was no witness, what would be the probability that the accident was caused by a yellow plate? **/2**

Answer: $P(C = y) = 5/100$

2. Your friend is working on a research project and has been given a small set of training data, but has not been given the test set. Instead your friend's supervisor keeps the test set, but allows the student to send models (with trained model weights) and receive the test error back. Your friend is very frustrated because he sent two sets of weights to be tested, weights $\mathbf{w}_A$ and $\mathbf{w}_B$. For model A, on the training set, your friend computed a mean-squared-error of 0.01, but received back an error of 0.67 from his supervisor. For model B, on the training set, your friend had an error of 0.71 and on the test 0.69. Your friend explains that for model A he used a penalty of $\lambda = 0.001$, and after receiving the test results, tried model B where he used $\lambda = 10$.

With this information, answer the following questions:

(a) Which model is overfitting and which model is underfitting? **/1**

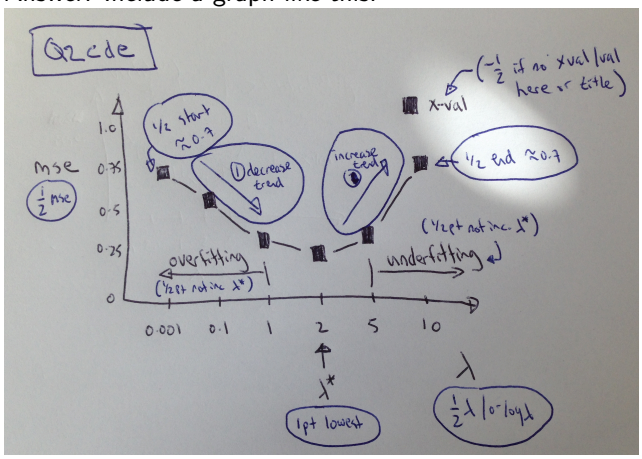
Answer: overfitting = model A; underfitting = model B

(b) You explain a procedure for selecting λ to your friend that will try to avoid overfitting and underfitting and only requires the training set. What is the procedure called? What is the algorithm? Why does it work? Be clear but brief. **/6**

Answer: The procedure is K-folds cross-validation (could leave out k-folds; could also use a validation set and not do k-folds). Algorithm: split your training set into K equally sized folds. For each value of λ, and for each fold, using the kth fold for test and the remaining as training data. Compute the cross-validation error (the total "test" error). Select λ with lowest error. Why does it work? It uses the held-out training set as unseen or test data, approximating the generalization performance, which is the primary goal for model selection.

(c) You apply the procedure to $\lambda \in \{0.001, 0.1, 1, 2, 5, 10\}$. Draw a graph that includes the error values in the question along with the results of the procedure (the values of $\log \lambda$ along the x-axis (equally spaced is ok), the values of the error along the y-axis). You can decide how to interpolate the error values, they just need to be a plausible outcome of the procedure relative to the error values in the problem statement. Label plots and axes accordingly. Remember you have run $\lambda \in \{0.001, 10\}$ on the train and test, but not on your procedure. **/4**

Answer: Include a graph like this:



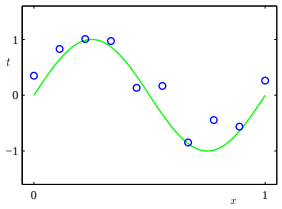
(d) Indicate on the graph which regions are overfitting and which are underfitting. **/1**

Answer: (on graph) the regions with $\lambda < \lambda^\star$ are overfitting and regions $\lambda > \lambda^\star$ are underfittting, though this could change depending on the values given.

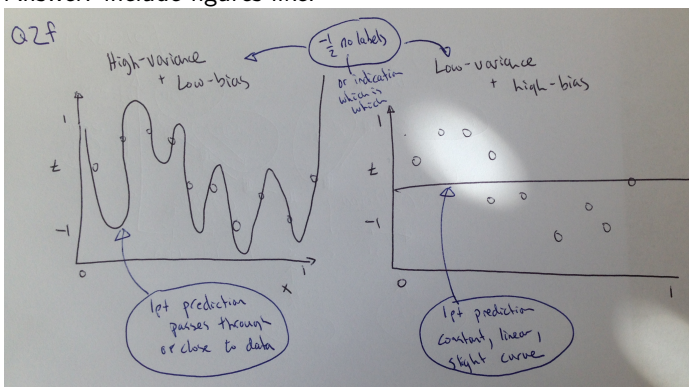
(e) Indicate on the graph the value of λ your friend should select. **/1**

Answer: (on graph) $\lambda^\star$ should be indicated somehow and it should have the lowest xval error.

(f) Your friend is still confused about what is going on. You explain the bias-variance error decomposition to him. Reproduce the figure below 2 times. In one, plot the solution $(y(x, \mathcal{D}))$ of a model with high-variance and low-bias and in the other plot, a model with low-variance and high-bias (both trained on the data set $\mathcal{D}$ shown as circles). Note the true regression function $h(\mathbf{x})$ is shown as a solid line. **/2**



Answer: include figures like:



Answer: high-variance and low-bias will pass through or close to the training data, but not smoothly; interpolation will give wild predictions. low-variance and high-bias will be a constant or line or slightly curving line.

(g) Consider the following error terms found in the expected loss:

i. $\int\{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t) d\mathbf{x} dt$

ii. $\int\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}, \mathcal{D})] - h(\mathbf{x})\}^2 p(\mathbf{x}) d\mathbf{x}$

iii. $\int\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}, \mathcal{D})] - y(\mathbf{x}, \mathcal{D})\}^2 p(\mathbf{x}) d\mathbf{x}$

Label the error terms as bias, variance, or noise. As modelers, which term(s) do we have control over? **/3**

Answer: i = noise; ii = bias; iii = variance; we have control over ii and iii.

3. Assume a classification problem with two classes $\mathcal{C}_0$ and $\mathcal{C}_1$. We observe the following data pairs: $\{t_n, x_n\} = \{(0, 1), (0, 1.5), (0, 2.0), (1, 2.5), (1, 3.0), (1, 3.5)\}$. Assume that if $t_n = 1$ the pair belongs to $\mathcal{C}_1$, otherwise to $\mathcal{C}_0$.

(a) Write down the prediction function $y_0(x, w_0, w_{00})$ and $y_1(x, w_1, w_{10})$ for a linear least-squares classifier. **/2**

Answer: $y_0(x, w_0, w_{00}) = w_0 \cdot x + w_{00}$ and $y_1(x, w_1, w_{10}) = w_1 \cdot x + w_{10}$

(b) Write down the prediction function $y(x, w, w_0)$ for a logistic-regression classifier. **/1**

Answer: $y(x, w, w_0) = 1/(1 + \exp(-w \cdot x - w_0))$

(c) What probability does the logistic-regression prediction function correspond to? **/1**

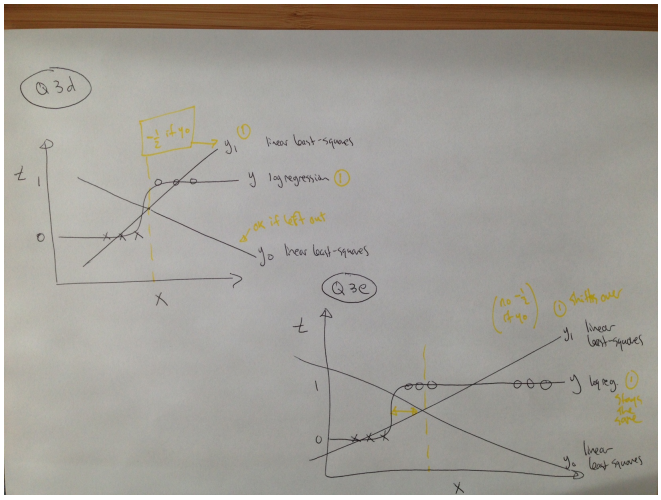
Answer: the class probability $P(C_1|x)$ or $P(t = 1|x)$.

(d) Make a graph of the data and plot the prediction functions for the two classifiers. **/2**

See next figure.

(e) Imagine you now receive 3 more data pairs: $\{(1, 10), (1, 11), (1, 12)\}$. Draw another graph including the new and original data and prediction functions for both classifiers based on all the data. **/2**

See figure.



(f) Explain why "too correct" data (the new pairs) affect the models differently by addressing the modeling assumptions and/or objective functions made by least-squares and logistic regression. **/4**

Answer: For least-squares, the objective function penalizes the linear function $y_1(x, w_1, w_{10})$ very strongly because it treats the targets as Gaussians. For even more correct data the penalty will be even more severe. For logistic regression, the solution doesn't change because the new targets are already correctly predicted and there is very little penalty in the cross-entropy term (the log term goes to 0). Although the main focus was the effect on least-squares, could mention the overfitting of logistic regression on separable data, though this would happen in both cases.

4. Consider the following general set-up. You have a data set of input-output pairs $\{\mathbf{t}, \mathbf{X}\}$, where $\mathbf{t}$ is an N by 1 vector of target values and $\mathbf{X}$ is an N by D matrix of input data. Assume that for model m there are parameters $\boldsymbol{\theta}_m$ and model hyper-parameters $\boldsymbol{\gamma}_m$. The likelihood function for the nth data pair is $p(t_n|\mathbf{x}_n, \boldsymbol{\theta}_m, \boldsymbol{\gamma}_m)$ and the prior distribution for $\boldsymbol{\theta}_m$ is $p(\boldsymbol{\theta}_m|\boldsymbol{\gamma}_m)$. E.g. if model m was the linear regression model studied in class, then $\boldsymbol{\gamma}_m = \{\alpha, \beta\}$ (the precisions of the prior and likelihood functions) and $\boldsymbol{\theta}_m = \mathbf{w}$, the regression weights. For the questions below, consider the general case, not the linear regression example.

(a) Write down the exact form of **maximum log-likelihood** learning for this model. Write your answer in the form $\boldsymbol{\theta}_m = \arg\max_{\boldsymbol{\theta}_m} O(\boldsymbol{\theta}_m, \boldsymbol{\gamma}_m, \mathbf{t}, \mathbf{X})$, but you fill in the details of $O(\boldsymbol{\theta}_m, \boldsymbol{\gamma}_m, \mathbf{t}, \mathbf{X})$ using the definitions in the problem statement. **/2**

Answer: $\boldsymbol{\theta}_m = \arg\max_{\boldsymbol{\theta}_m} \sum_n \log p(t_n|\mathbf{x}_n, \boldsymbol{\theta}_m, \boldsymbol{\gamma}_m)$ (using vector forms ok too)

(b) Do the same for **maximum a-posteriori log-likelihood** learning. **/2**

Answer: $\boldsymbol{\theta}_m = \arg\max_{\boldsymbol{\theta}_m} \sum_n \log p(t_n|\mathbf{x}_n, \boldsymbol{\theta}_m, \boldsymbol{\gamma}_m) + \log p(\boldsymbol{\theta}_m|\boldsymbol{\gamma}_m)$

(c) Write down the expression for the **evidence** for model m using the product and sum rule. **/2**

Answer: $p(\mathbf{t}|\mathbf{X}, \boldsymbol{\gamma}_m) = \int p(\mathbf{t}, \mathbf{X}, \boldsymbol{\theta}_m, \boldsymbol{\gamma}_m) p(\boldsymbol{\theta}_m|\boldsymbol{\gamma}_m) d\boldsymbol{\theta}_m$

(d) Write down the expression for the **posterior distribution** of $\boldsymbol{\theta}_m$, using the general probability densities defined in the problem statement. Label the likelihood, prior, evidence terms. Ensure that all the conditioning statements are correct. **/3**

Answer: $$p(\boldsymbol{\theta}_m|\mathbf{t}, \mathbf{X}, \boldsymbol{\gamma}_m) = \frac{\overbrace{p(\boldsymbol{\theta}_m|\boldsymbol{\gamma}_m)}^{\text{prior}} \overbrace{p(\mathbf{t}|\mathbf{X}, \boldsymbol{\theta}_m, \boldsymbol{\gamma}_m)}^{\text{likelihood}}}{\underbrace{\int p(\mathbf{t}, \mathbf{X}, \boldsymbol{\theta}_m, \boldsymbol{\gamma}_m) p(\boldsymbol{\theta}_m|\boldsymbol{\gamma}_m) d\boldsymbol{\theta}_m}_{\text{evidence}}}$$

(e) Describe one way that the posterior distribution can be used to make predictions for new input vectors $\mathbf{x}^\star$. You can use words or write the expression. **/3**

Answer:
either: $p(t^\star|\mathbf{x}_\star, \mathbf{t}, \mathbf{X}\boldsymbol{\gamma}_m) = \int p(t^\star|\mathbf{x}_\star, \boldsymbol{\theta}_m) p(\boldsymbol{\theta}_m|\mathbf{t}, \mathbf{X}, \gamma_m) d\boldsymbol{\theta}_m$
or:
$p(t^\star|\mathbf{x}_\star, \mathbf{t}, \mathbf{X}\boldsymbol{\gamma}_m) = \frac{1}{S} \sum p(t^\star|\mathbf{x}_\star, \boldsymbol{\theta}_m^{(s)}) \qquad \boldsymbol{\theta}_m^{(s)} \sim p(\boldsymbol{\theta}_m|\mathbf{t}, \mathbf{X}, \boldsymbol{\gamma}_m)$
or use MAP estimate from posterior for predictions (technically correct, but not a good use of a posterior distribution.)

(f) Assume that there is an analytic solution to the evidence computation in part (c) above, for both model m and also for another model s with hyperparameters $\boldsymbol{\gamma}_s$. How can we use these analytic solutions to select the best model? **/2**

Answer: Compute the two evidences $p(\mathbf{t}|\mathbf{X}, \boldsymbol{\gamma}_m)$ and $p(\mathbf{t}|\mathbf{X}, \boldsymbol{\gamma}_s)$; select model with largest evidence. Could also mention model comparison, averaging.

5. Imagine you have written some computer vision software for a flying drone. Your software will predict—10 times per second—whether the drone will hit a tree in the next second or whether there is no tree. In other words, can compute $P(C_0 = \text{tree}|\mathbf{x})$ (which implies $P(C_1 = \text{no tree}|\mathbf{x}) = 1 - P(C_0 = \text{tree}|\mathbf{x})$). The action associated with predicting "tree" is to quickly move (swerve) perpendicular to the drone's current direction (a_0: action="swerve"). The action associated with predicting "no tree" is to continue following the current direction (a_1: action = "continue"). You estimate a loss of 100 if the drone hits a tree, and a loss of 1 every time the drone unnecessarily swerves to avoid a non-existent tree. With this information answer the following questions:

(a) Write down the loss matrix associated with this problem. Make sure the rows and columns are labeled. **/3**

Answer:

	tree (swerve)	not tree (continue)
tree	0	100
not tree	1	0

(b) At one moment the drone predicts $P(C_0|\mathbf{x}) = 0.15$. The drone needs to make a decision. Compute the expected losses for each possible decision/action. What action will the drone take? **/5**

Answer: compute both expected losses and pick action with lowest expected loss.

$$\begin{aligned} R(a_0) &= L_{00}P(C_0|\mathbf{x}) + L_{10}P(C_1\mathbf{x}) \\ &= 1 * 0.85 = 0.85 \\ R(a_1) &= L_{01}P(C_0|\mathbf{x}) + L_{11}P(C_1\mathbf{x}) \\ &= 100 * 0.15 = 15 \end{aligned}$$

Since $R(a_0) < R(a_1)$, choose a_0=swerve.

(c) What value of the prediction $P(C_0|\mathbf{x})$ will cause the drone to be unable to make a decision? **/2**

Answer: equate the two expected losses:

$$\begin{aligned} 1 - P(C_0|\mathbf{x}) &= 100P(C_0|\mathbf{x}) \\ p(C_0|\mathbf{x}) &= 1/101 \end{aligned}$$