Problem 1

(A) A hypothesis class is the set from which hypotheses are chosen. The hypothesis class of a linear model are all hyperplanes, which are usually given of the form $w^\intercal x + b$.

(B)

(C) In supervised learning, for each input point x, you are given its corresponding and correct output y. On the other hand, in unsupervised learning, for each input point x, we do not get any output whatsoever.

(D) The training data is what the machine learning algorithm is trained on to develop a hypothesis; the test data is what the hypothesis is tested on to see how well it performs.

(E) The assumption is that the training data is sampled according to the probability distribution on the universe.

(F) For classification, hinge loss is good because once you're right, the error goes to zero (eventually), so the hypothesis doesn't have to worry to much about getting things exactly right. For real valued functions, squared loss is good because it always gives an accurate representation of how far away you are from the desired value.

Problem 2

(A) Minimizing the negative log-likelihood looks like We can do this

$$\text{argmax}_w \ln\Big(\prod_{(x_i,y_i)\in S} P(y_i|x_i,w)\Big)$$

because the quantity inside is nonnegative (probabilities are nonnnegative) and natural log is monotonically increasing so if you maximize it you maximize its argument. Then clearly we can just minimize the negative to get the following.

$$\text{argmin}_w - \ln\Big(\prod_{(x_i,y_i)\in S} P(y_i|x_i,w)\Big)$$

And then, applying log rules, we get

$$\text{argmin}_w \sum_i - \ln\Big(P(y_i|x_i,w)\Big)$$

(B) It is typically easier to solve the negative log-likelihood formulation because you can just use gradient descent.

(C) We can let

$$P(y|x,w) = \frac{1}{1+e^{-yw^\intercal x}}$$

$$w_{t+1} \leftarrow w_t - \nabla \sum_i - \ln\Big(\frac{1}{1+e^{-y_i w_t^\intercal x_i}}\Big)$$

$$w_{t+1} \leftarrow w_t - \nabla \sum_i \ln\Big(1+e^{-y_i w_t^\intercal x_i}\Big)$$

$$w_{t+1} \leftarrow w_t - \sum_i \frac{-y_i x_i}{1+e^{y_i w_t^\intercal x_i}}$$

Problem 3

(A) There are K training sessions

(B) The training and validation sets are of size N/K.

(C) Training sets are allowed to overlap during different sessions, but validations sets are not.

(D) We do cross validation to use a large training set (which gives us a better hypothesis) while simultaneously using, in effect, a large testing set (which allows use to accurately test how good that hypothesis is).

(E) The largest possible value of K is $K = N$ (you could always just randomly choose a hypothesis and train on nothing!).

Problem 4

(A)

$$
\begin{aligned}
&E_S\Big[(h_S(x) - y)^2\Big] \\
&= E_S\Big[(h_S(x) - \bar{h} + \bar{h} - y)^2\Big] \\
&= E_S\Big[(h_S(x) - \bar{h})^2 + (\bar{h} - y)^2 + 2(h_S(x) - \bar{h})(\bar{h} - y)\Big] \\
&= E_S\Big[(h_S(x) - \bar{h})^2\Big] + E_S\Big[(\bar{h} - y)^2\Big] + E_S\Big[(2(h_S(x) - \bar{h})(\bar{h} - y)\Big] \\
&= E_S\Big[(h_S(x) - \bar{h})^2\Big] + (\bar{h} - y)^2 + 0 \\
&= E_S\Big[(h_S(x) - \bar{h})^2\Big] + (\bar{h} - y)^2
\end{aligned}
$$

(B) Bias is called bias because it represents how close/far the average hypothesis is to the actual model. So if bias is high, then errors are likely to be high as well. Variance is called variance because it represents how close each hypothesis is from the average, how variant the model class is.

(C) If the complexity of the model increases, then variance will go up, since there will be a large variety of models to choose from. Bias, however, will go down, for the same reason; given the larger variety of models to choose from, it should be easier to fit the data.

Problem 5

We can use

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)}$$

(A)

$$
\begin{aligned}
P(h_1|E) &= \frac{.75 \times .5}{.5} = .75 \\
P(h_2|E) &= \frac{.5 \times .25}{.5} = .25 \\
P(h_3|E) &= \frac{1 \times .25}{.5} = .5
\end{aligned}
$$

(B)

$$
\begin{aligned}
P(h_1|E) &= \frac{.25 \times .75}{.5} = .375 \\
P(h_2|E) &= \frac{.25 \times .25}{.5} = .125 \\
P(h_3|E) &= \frac{.5 \times .5}{.5} = .5
\end{aligned}
$$

(C)

$$P(h_1|E) = \frac{.25 \times .375}{.5} = .1875$$
$$P(h_2|E) = \frac{.75 \times .125}{.5} = .1875$$
$$P(h_3|E) = \frac{0 \times .5}{.5} = 0$$