CS434 Machine Learning and Data Mining Midterm

Study online at https://quizlet.com/_9svg8y

What occurs if our model makes no assumptions?	No learning occurs!
How many possible mappings exist for a binary vector of length d mapping to a binary output?	2^(2^(d)) possible mappings
What if we have n data points?	Still 2^(2(^(d) - n) -- SO many!
What assumption do we make for kNN?	(i) Label changes smoothly as features change in local regions; (ii) Each feature affects the output independently
What assumption do we make for logistic regression?	(i) The relationship between input and output can be expressed linearly; (ii) Label changes smoothly as features change locally; (iii) Each feature effects output independently
What algorithms share a similar assumption?	SVM, perceptron, and linear regression - examples can be linearly separated or predicted with a linear model
What is modelling error, and how do we reduce it?	You chose the wrong model / hypothesis space. Reduce by choosing a better model.
What is estimation error, and how do we reduce it?	You didn't have enough data. Reduce by adding more data (infinite data).
What is optimization error, and how do we reduce it?	Your model was not optimized well. Reduce by optimizing longer (infinite training time), by substituting with a better optimization algorithm or model, or applying more expensive optimization.
What is Bayes' error, and how do we reduce it?	Your model was unable to distinguish between overlapping distributions. Irreducible with a given dataset, UNLESS a new feature is introduced which meaningfully discriminates between instances with the same features but different label. If impossible, this error is called "irreducible".
What is overfitting?	Model performs well on training, but poorly on validation or test data.
What is underfitting?	Model performs badly on training, validation, and test data.
What is model selection?	The process of finding the proper hypothesis space (AKA the "model class") that neither underfits nor overfits. This is challenging!
What is kNN?	A type of model that predicts the label of an unknown example by measuring some weighted average of the k neighbors around it that are closest in distance (k-nearest). In the vanilla model, all weights w_i are = 1.
What are the effects of extreme k-values in k-NN?	At k=1, training error is zero when Bayes error is zero. At k=n, every point is a neighbor, leading to the majority class being predicted everywhere in the dataset.
What would a generalization of k-NN to a regression task entail?	Take the weighted average of all the neighbors, and predict that value. For 1-NN, just take the value of the closest neighbor.
For a regression k-NN, what would we predict when k=n?	The predicted value would equal the average value of the dataset.
What are the problems with k-NN?	- Computationally expensive: requires O(nd) for every test point, although proper choice of data structure can reduce this cost. - Massive datasets require lots of examples, but this can be reduced if we remove "unimportant" examples lying in the "safe" region with many of the same labels. - The relative scale of features matters as well, because large distances between large features matters more than small distances between small features, so we should scale features [0,1]. - Irrelevant features also contribute to distance, meaning that distances matter less (in high-dims, everything is the same distance), so we should remove unnecessary (unimportant) features (that are not predictive of the output).

What hyperparameters can we tweak for k-NN?	(i) Distance metric;
	(ii) k;

	(iii) Output function (weighting vs not); (iv) Weighting function, and bandwidth/variance
How big should the validation set be?	The larger the validation set, the more reliable our model selection choices will be in terms of generalization to test set.
How should we deal with a small dataset for k-NN? How do we estimate validation performance?	Run cross-validation! This means making K-many (unrelated) separate validation sets out of different subsets of (1/K)th of the data, then train on the other (K-1) subsets. To estimate validation performance we take the mean of the accuracies of the subsets during K-fold cross validation.
After picking best hyperparameters and evaluating on test data, should you tune again?	No! Your model is no longer useful!
How does LOO cross-validation work?	When K=n, we treat only ONE example in each validation split, so there are n-many splits with one training example in each. That way, we leave out ONLY the one we're looking for, then select whichever example is closest to the one we're looking for (that's not it!). Essentially, take the label of the one that's closer.
What is the goal of a Maximum Likelihood Estimation? (MLE)	To find parameters which make the data we OBSERVE empirically to be the most likely.
How do we perform MLE?	1.) Assume a probabilistic model for how the data was generated, w.r.t. some parameters ¸. 2.)Find ¸_MLE which maximizes the probability (likelihood) to generate the training data under the probabilistic model.
Why is MLE nice or useful?	- Frequently produces intuitive parameter estimates. - MLE is optimal if the model class is correct (e.g. Normal model for data that is actually normally distributed) - unoptimal if our model is wrong.
How do we perform MLE?	1.) Collect our dataset. 2.) Make our model assumption about the distribution of our random variable X. 3.) Write out likelihood of training data as a function of parameters ¸. 4.)Find ¸_MLE = argmax(L(¸))by taking derivative of log likelihood and setting equal to zero.
How does a prior belief effect our expectation of future measurements?	It smoothes out our data to represent more our belief about how the probability should unfold rather than what we measure. This is most useful at low values of elements in the dataset.
Compute Maximum A Priori parameter ¸for a coin flip with one head.	When we have one head flip, our likelihood function gives zero likelihood to tails and one to heads, manifesting as a straight line with slope 1, intercept zero. By adding our Prior = Beta(2,2), we expect it to look more like a (parabolic) beta distribution with average ¸_prior = 0.5. Our posterior, after considering our prior, is ¸_MAP = (3-1)/(3+2-2) = (2/3)
What happens to ¸_MAP as number of observed data points gets larger?	The fake heads/tails from the prior matter less, so our distribution looks more like the dataset actually is.
What is the difference between a MLE and a MAP estimate?	MAP maximizes the posterior (i.e., the likelihood times the prior (L(¸)*Pr)) whereas MLE maximizes the likelihood (L).
What is the equation for sum of squared error? (SSE(w))	SSE(w) = sum_{i=1}^{n} (y_i - w x_i)^2 = (y-Xw)^T @ (y-Xw)
Any downsides to linear regression?	Yes, it's susceptible to outliers, because we sum SQUARED errors, rather than just abs(err). We say this is because there is very low density in the tails of Gaussian distributions.
How do we perform linear regression using matrix algebra?	Take deriv of SSE and set = 0. Solve for: w* = (X^T @ X)^-1 X^T y
What is our generative story for a dataset?	"The true distribution matches a given model (linear, quadratic, sinusoidal, etc.) with some normally-distributed error function added to the true distribution."
How do we maximize the log-likelihood w.r.t. w?	Find w* s.t. SSE is as small as possible! Thus Linear regression is just MLE of a linear model, but with Gaussian noise added!

How do we fit a nonlinear dataset linearly?	Solve a linear regression problem in a feature space that is non-linear in the original input. E.g., add x^2 column. This is called our "basis function".
When can we use linear regression over basis functions to find weights?	Whenever the weights are linear with the function of x. If they're NOT linear, transform them! (with logs, exponents, etc.)
What is a regularizer?	A term added to the weight vector to simplify a model. Results in $$w* = argmin_w (L(w)) + » *reg(w)$$
What does the »term in a regularizer do?	Larger »leads to simpler models and less overfitting. Smaller »leads to a more complex model but improve the fit.
What is the regularizer based on?	Often, the norm of the weight vector.
What does the sigmoid function map?	(-infty, infty) -> (0,1)
What is logistic regression?	The construction of a model to determine which of two classes a new data point belongs to. Uses the logistic function to determine hoq likely it is that the point is class 0 versus class 1.
What is the logistic regression model assumption?	P(y_i=1 | x_i, w) = Ãw(^T x_i); AND P(y_i=0 | x_i, w) = 1 - Ãw(^T x_i)
What is the decision boundary for logistic regression?	Just the line w^T x = 0
What must we use for logistic regression to determine MLE?	Non-analytical --> Gradient descent (iterative)
What is similar between logreg and perceptron?	Both are linear classifiers
What is different between logreg and perceptron?	Logreg uses sigmoid, giving a probabilistic interpretation, but must be solved using gradient descent or other optimization methods. Perceptron uses the "sign" function with no probabilistic interpretation, and is trained using the "perceptron" algorithm.
What is the perceptron learning algorithm?	Initialize weights randomly. For each example (x_i, y_i) in D, until no errors or maxiters, If (y_i * w^T x_i < 0) (if misclassified), w = w + ±*y_i*x_i
What are limitations of the perceptron learning algorithm?	- Arrives at different solutions based on initial random weight vector. - Correcting for a misclassification moves the decision boundary --> misclassify previous correct examples. Thus, it must go over training examples multiple times (epochs). - Terminates if w not updated during an epoch. - While guaranteed to converge for LINEARLY SEPARABLE data, NON-linsep data fails to converge and may be arbitratily bad if reaches maxiter
What does independence mean?	P(X,Y) = P(X)P(Y); P(X|Y) = P(X); P(Y|X) = P(Y)
What does conditional independence mean?	P(X,Y|Z) = P(X|Z)P(Y|Z); P(X|Y,Z) = P(X|Z); P(Y|X,Z) = P(Y|Z)
What is the takeaway from conditonal independence?	Events that are dependent in general can be made independent given some other observation (data).
What is a discriminative classifier?	Learns P(y|x) directly. Ex: logistic regression. Classifies according to argmax P(y|x)
What is generative classifier?	Learns P(x,Y) and P(y), then computes P(y|x) using Bayes' rule. Ex: Naive Bayes. Classifies according to argmax P(y|x)
What is the naive Bayes assumption?	Each feature is conditionally independent given the class label: P(x|y) = prod_{i=1}^{d} P(x_i|y)
What is super convenient about the Naive Bayes assumption?	Reduces the parameter cost of learning P(x|y) from c*(m-1)^d to c*(m-1)*d (For P(x_i|y)) + (c-1) (for class prior P(y))
Steps of Naive Bayes model?	1.) Learn the conditional P(x_i|y=c) for each feature x_i and class c. 2.) Estimate P(y=c) as a fraction of records with y=c for each class c. 3.) For a new example x = [x_1, ... x_m]^T, predict argmax_{c=1,2,3,...,k} P(y=c) * prod_{i=1}^{d} P(x_i | y=c)
What are problems with Naive Bayes?	The zero-probability problem makes the whole product zero if one probability is zero--i.e. if a new email contains a new word we've never seen in the training emails for our spam filter model. To fix this, we can apply Laplace smoothing for binary variables.

CS434 Machine Learning and Data Mining Midterm

Study online at https://quizlet.com/_9svg8y

What is Laplace smoothing?	For a binary variable x_i, add a small prior to p(x_i|y=c). Often, a Beta(1,1) prior (added to the estimated conditional distribution) is sufficient.
How is a Hard-margin SVM different from a soft-margin one?	Soft-margin has slack penalty which allows some violations (up to a limit) to occur when making the margin as wide as possible.
How is a Hard-margin SVM the same as a soft-margin one?	Given optimal alphas, the weight vectors are the same: w* = sum_{i=1}^{n} ±i^{*} * y_i * x_i
What happens when slack penalty C goes to zero?	Errors are NOT penalized, letting the margin grow infinitely wide.
What happens when slack penalty C goes to infinty?	Soft-margin -> hard margin, meaning errors are very heavily penalized (infinitely so...)
What is a kernel function?	Those functions that satisfy: º(a,b) = Æ(a)*Æ(b) for some mapping function Æ. Essentially computes similarity between a and b.
How many functions have kernels?	NOT all of them, but a lot of popular ones do!
What is the kernel trick?	Computing similarity in the high-dim mapping space without transforming the data into that space, using a kernel function.
What is the summation closure property of kernel functions?	º(a,b) = º_1(a,b) + º_2(a,b)
What is the constant multiple closure property of kernels?	º(a,b) = c*º_1(a,b) for c>0
What is the product closure property of kernels?	º(a,b) = º_1(a,b)*º_2(a,b)
How to use kernels to make SVM better?	Just replace the dot product with a kernel between those two feature vectors!