Describe machine learning; A form of applied statistics with

- increased emphasis on the use of computers to statistically estimate complicated functions and
- a decreased emphasis on proving confidence intervals around these functions

(subjective?)

Define learning for a program (Mitchell 1997); A computer program is said to learn from experience E wrt some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Describe supervised learning;

- Experience = dataset (random vectors $\mathbf{x}$) containing features, where each example is associated with a label or target $\mathbf{y}$.
- Learn to predict $\mathbf{y}$ from $\mathbf{x}$, usually by estimating $p(\mathbf{y}|\mathbf{x})$.

Note not completely distinct from unsupervised learning (e.g. infer conditionals from joint $p(x)$ which is learned in unsupervised learning).

Described unsupervised learning;

- Experience = dataset (examples of random vector $\mathbf{x}$) containing many features
- Learn useful properties of the structure of the dataset, usually attempt to implicitly or explicitly learn probability distribution $p(\mathbf{x})$.

Note not completely distinct from supervised learning (e.g. by decomposing joint into conditionals).

Briefly describe reinforcement learning (contrasted against supervised or unsupervised learning); Dataset is not fixed. RL algorithms interact with an environment, so that there is a feedback loop between the learning system and its experiences.

Design matrix; Way of describing a dataset. Each row is a different example, each column a different feature.
but data vectors may not all be of the same size → can describe using a set.

Linear regression example;

Normal equations; Given a matrix equation $\mathbf{Ax} = \mathbf{b}$, the normal equation is that which minimises the sum of square differences between the left and right sides of $A^TAx = A^Tb$.
Called normal eqn because $(b - Ax)$ is normal to the range of A.

Linear vs affine function; Affine = linear function with translation, e.g. $wx + b$, linear function of x is just wx.

Why is the intercept term in linear regression called the bias?; Idea: output is biased towards this qty absent of any input.
Not the same as statistical bias (of an estimator, expected val not equal to true val).

1 Capacity, Overfitting and Underfitting

What separates ML from optimisation?; We want the generalisation error to be low as well (vs just the training error).

Generalisation; ability to perform well on previously unobserved inputs

Generalisation error; Expected value of error from new input (drawn from dist of inputs we expect the system to encounter in practice).

iid assumptions (train and test sets); Assume (1) examples in each dataset indep from each other, (2) train and test identically distributed (drawn from same prob dist, i.e. data-generating dist, as each other).

Underfitting; model not able to obtain a sufficiently low error on the training set

Overfitting; gap between training error and test error is too large

Capacity (of a model); ability to fit a wide variety of functions. (Low: may underfit, High: may overfit by memorising properties of training set that do not serve models well on test set).
Can control by e.g. choosing hypothesis space of model.

Hypothesis space (of a model); Set of functions the learning algorithm is allowed to select as being the solution.

When there are more parameters than training examples; there may be infinitely many functions that pass exactly through the training points.
Have little chance of choosing a solution that generalises well when so many wildly different solutions exist.

Representational capacity;

- Family of functions the learning algorithm can choose from when varying the parameters in order to reduce a training objective.
- vs effective capacity: take into account alg does not actually find best fn, but merely one that sig reduces training error

(TODO: clarify?)

Effective capacity; Capacity taking into account imperfections of optimisation algorithm (e.g. choose parameters within model family that min training error instead of choosing 'best function') (TODO: clarify?)

Occam's Razor; Simplicity. Among competing hypotheses that explain known observations equally well, we should choose the 'simplest' one.

How to quantify model capacity; e.g. is using VC dim.

Vapnik-Chervonenkis dimension (VC dim);

- Measures the capacity of a binary classifier.
- Largest possible value of m for which there exists a training set of m different $\mathbf{x}$ points that the classifier can label arbitrarily.

Statistical learning theory results on discrepancy between train and test error;

- Bounded from above by qty that grows as model capacity grows (higher capacity, larger possible discrepancy)
- but shrinks as the number of training examples increases (more training examples, smaller possible max discrepancy)
- But rarely used in practice bc bounds quite loose, also hard to determine capacity of DL algorithms.
- Especially hard in DL because effective capacity limited by capabilities of optimisation algorithm, and we have little theoretical understanding of general nonconvex optimisation problems involved in deep learning

Vapnik and Chervonenkis, Vapnik, Blumer et al.

Typical shape of test error as a function of model capacity; U-shaped (can describe as underfitting regime/zone, optimal capacity, overfitting regime/zone)

Nonparametric models (describe wrt capacity, give e.g.s); can have arbitrarily high capacity since learned fns are not described by a finite parameter vector. E.g.s

- Completely nonparametric:
  - Abstract: alg that searches over all possible P(X) dists
  - Practical: nearest neighbour regression
- Partially parametric: wrap parametric inside another alg that increases num of params as needed.

Can a model with less than optimal capacity achieve the Bayes error?; No. Any fixed parametric model with less than optimal capacity will asymptote to an error value that exceeds the Bayes error.

Can a model have optimal capacity and still have a large gap between training and test errors?; Yes. May be able to reduce gap by gathering more training examples.

Bayes error; Error incurred by an oracle making predictions from the true distribution $p(\mathbf{x}, y)$.

Nearest neighbour regression; Looks up nearest entry x_i in training set and returns y_i.

No Free Lunch theorem;

- Averaged over all possible data-generating distributions, every classification algorithm has the same error rate when classifying previously unobserved data points. (Wolpert, 1996)
- So goal is not to seek 'universally best learning alg', but to understand what dists are relevant to real-world tasks and what algs perform well on data drawn from dists we care about.

Can expected test error increase if the number of training examples increases? No.

Regularisation; Any modification we make to a learning alg that is intended to reduce its test error but not its training error.

- Way of expressing a preference for certain solns in the fn's hypothesis space.

Examples of modifying learning algorithm;

- Exclude functions from hypothesis space (infinitely strong preference against fn)
- Express preferences for one soln over another in alg's hypothesis space (e.g. by including regulariser)

Weight decay; $\min J(\mathbf{text}) = \min MSE_{train} + \lambda \mathbf{w}^{\mathbf{T}}\mathbf{w}$, latter term is λ x regulariser $\Omega(\mathbf{w})$.

Regulariser; Penalty added to cost fn to regularise a model

2 Hyperparameters and Validation Sets

Hyperparameters; Params (control alg behaviour) that are not adapted by the learning algorithm itself.

- Often bc not appropriate to learn hyperparam on training set, e.g. model capacity, since will just choose max val and overfit.
- Can use validation set to learn hyperparameters

Validation set;

- Guides selection of hyperparameters
- Can use validation error to estimate test error

What are the disadvantages to having the same ML benchmark for a long time?; May become stale, overly optimistic re performance on benchmark.

Motivation for cross-validation;

- If dataset is small, there is statistical uncertainty around estimate of average test / validation error, so it's hard to claim alg A is b etter than alg B on a particular task.
- So use all examples in estimate of mean test error, at price of increasing computational cost.

Cross-validation;

- If dataset is small, there is statistical uncertainty around estimate of average test / validation error, so it's hard to claim alg A is b etter than alg B on a particular task.

- So in CV, use all examples in estimate of mean test error, at price of increasing computational cost.
- K-fold CV: partition of data formed by partitioning data into k subsets, esti test error by taking avg test error across t trials.
- Problem: no unbiased estimates of the variance of such average error estimates exist, but approximations are usually used.

Problem with k-fold CV; (1) No unbiased estimates of the variance of average test error est (across the k folds), but approximations are usually used. Also (2) increases computational cost.

3 Estimators, Bias and Variance

Bias (statistical); measures expected deviation from the true value of the function or parameter. $E[\hat{\theta_m} - \theta]$.

Variance;

Variance of an estimator provides; A measure of how we'd expect the estiamte we compute from data to vary as we independently resample the dataset from the underlying data-generating process.

Does the square root of the sample variance or the square root of the unbiased estimator of the variance provide an unbiased estimate of the standard deviation?; Neither - both approaches tend to underestimate the true stdev. But they are still used in practice. Sqrt of unbiased estimator of the variance (1/m+1) is less of an underestimate.

Using stdev in ML; Common to say alg A is better than alg B if upper bound of 95% CI for error of alg A is lower than lower bound of 95% CI for error of alg B.
Test error often calc as sample mean of error on test set, so using CLT, can say approx dist Gaussian, so CI is sample mean +/- 1.96 stdev.

Bias-variance tradeoff; As inc model capacity, bias tends to decrease and variance tends to increase.
Note MSE= $Bias^2 + Var$. Min MSE is a kind of compromise.

How does increasing model capacity tend to affect bias and variance?; Tends to increase varaince and decrease bias.

Consistency; $p \lim_{m \to \infty} \hat{\theta}_m = \theta$.
That is, $\forall \epsilon > 0, P(|\hat{\theta}_m - \theta| > \epsilon) \to 0$ as $m \to \infty$.
Convergence in probability, 'weak consistency'.

Strong consistency; Almost sure convergence of $\hat{\theta}$ to θ.
i.e. $p(\lim_{m \to \infty} \mathbf{x}^{(m)} = \mathbf{x}) = 1$. for sequence of random variables $\mathbf{x}^{(i)}$.

Does consistency imply asymptotic unbiasedness?; Yes.

Does asymptotic unbiasedness imply consistency?; No. E.g. first sample of dataset is unbiased estimator of mean, but is not the case that $\hat{\theta} \to \theta$ as $m \to \infty$.

4 Maximum Likelihood Estimation

Describe maximum likelihood estimation; $\theta_{ML} = \arg\max_\theta p_{model}(\mathcal{X};\theta)$, where $\mathcal{X}$ is a set of examples drawn from the underlying data-generating distribution $p_{data}(\mathbf{x})$. Equivalently, maximise log-likelihood (to prevent numerical underflow or overflow from products).

Relationship between maximum likelihood and the KL divergence; Maximising expected log-likelihood is equivalent to minimising the KL divergence (with respect to parameters θ). Helpful since KL has known minimum of zero, whereas NLL can become negative when $\mathbf{x}$ is real-valued.

We can reformulate maximising likelihood as minimising negative log-likelihood. Can the NLL become negative?; Yes, if $\mathbf{x}$ is real-valued.

KL divergence; $D_{KL}(\hat{p}_{data}||p_{model}) = E_{\mathbf{x}\sim\hat{p}_{data}}[\log\hat{p}_{data}(\mathbf{x}) - \log p_{model}(\mathbf{x})]$

Cross-entropy; Any loss consisting of a negative log-likelihood is a cross-entropy between (1) $\hat{p}_{data}$, the empirical distribution defined by the training set and (2) p_{model}, the probability dist defined by the model.
i.e. not just NLL of Bernouille or softmax dist.

How can we derive linear regression?; Derive MSE criterion using maximum likelihood on a linear Gaussian (noise) model.

Conditional maximum likelihood estimator $\theta_{ML} = \arg\max_\theta p_{model}(Y|X;\theta)$

5 Bayesian Statistics

Prior;

6 Supervised Learning Algorithms

Logistic regression;

Support vector machines;

Kernel trick;

Gaussian kernel;

Radial basis function kernel;

Template-matching;

Kernel machines or kernel methods;

Support vectors;

K-nearest neighbours;

Decision trees;

7 Unsupervised Learning Algorithms

Principal Components Analysis;

K-means clustering;

8 Stochastic Gradient Descent

Stochastic Gradient Descent;
    Minibatch;

9 Challenges Motivating Deep Learning

Curse of Dimensionality;
    Smoothness Prior / Local Constancy Prior;
    Local kernels;
    Manifold;
    Manifold learning;
    Manifold hypothesis;