Mid-term exam info

- In-class exam
- Closed books/notes, closed electronics, 2 sided A4 hand-written (not printed) cheat sheet allowed (upload somewhere by some date)
- Videos on
- Academic integrity violations severe consequences

- Multiple choice
- Via Gradescope or Canvas
- ~25 questions – 80 mins (3-4 mins/question)
  - ~40% are hard (think open book), ~60% are easy
- Grading will be curved
- Ask questions via private Zoom chat

Topics

- Basics - Probability, Matrix/vector calculus, Optimization (convexity etc)
- Basic ML concepts – training vs test data, overfitting, generalization, ML tasks, loss metrics, optimal classifier/regressor, decision boundaries
- Distribution/Density estimation – MLE, MAP, Histogram, Kernel density estimation
- Classification – Naïve Bayes, Logistic Regression, Neural Networks, k-Nearest Neighbors
- Regression – Linear Regression, Ridge, Lasso, Neural Networks, Kernel regression

Comparison chart (classification)

Algorithm	Generative/ Discriminative	Assumptions	Decision boundary	Loss function	Training
Naïve Bayes	G	$P(X\|Y)$, $P(Y)$ — conditional ind; Ber/Multin; Gaussian	GNB:		
Logistic Regression	D	$P(Y=1\|X) = \frac{e^{\sum w_i x_i}}{1+e^{\sum w_i x_i}}$	$e^{\sum w_i x_i} \geq 1$		
Neural Networks	D				
k-Nearest Neighbors	D	$P(Y\|X) \approx \frac{K_y}{K}$			

Comparison (regression)

Algorithm	Generative/ Discriminative	Assumptions	Decision boundary	Loss function	Training
Linear Regression (Lasso, Ridge) → l_1 & l_2					
Neural Networks					
Kernel regression					

Practice problems (basics)

- Which of the following expressions is equivalent to P(A,B|C)?

(a) $\frac{P(A,B)}{P(C)}$

(b) $\mathbf{P(A|C)P(B|A,C)}$

(c) $P(C|A,B)P(B|A)P(A)$

(d) $\frac{P(A)P(B|A)P(C|A,B)}{\sum_{a\in A} P(A=a)P(C|A=a)+\sum_{b\in B} P(B=b)P(C|B=b)}$

- The function x^TAx where x is a d-dim vector and A is a dxd matrix is
  - Convex if A is rank 1
  - Convex because its quadratic in x
  - Concave if A has all negative eigenvalues

- Stochastic Gradient descent is faster but uses more memory than regular gradient descent.

Practice problems (ML intro)

- When the feature space is larger, overfitting is less likely.

- Which of the following are supervised learning tasks:
  - Predicting the proportion of students in class who slept < 7 hours
  - Predicting rating of a movie given movie genre
  - Tagging tweets that are racially provocative

$\equiv P(Y)$

- A classifier that attains 100% accuracy on the training set and 70% accuracy on test set is better than a classifier that attains 70% accuracy on the training set and 75% accuracy on test set.

Practice problems (density/distribution estimation)

- You have received a shiny new coin and want to estimate the probability θ that it will come up heads if you flip it. A priori you assume that the most probable value of θ is 0.5. You then flip the coin 3 times, and it comes up heads twice. Which will be higher, your maximum likelihood estimate (MLE) of θ, or your maximum a posteriori probability (MAP) estimate of θ?.

- The maximum likelihood estimate of model parameter α for the random variable $y \sim N(\alpha x_1 x_2^3, \sigma^2)$, where x_1 and x_2 are random variables, can be learned using linear regression on n iid samples of (x_1, x_2, y)

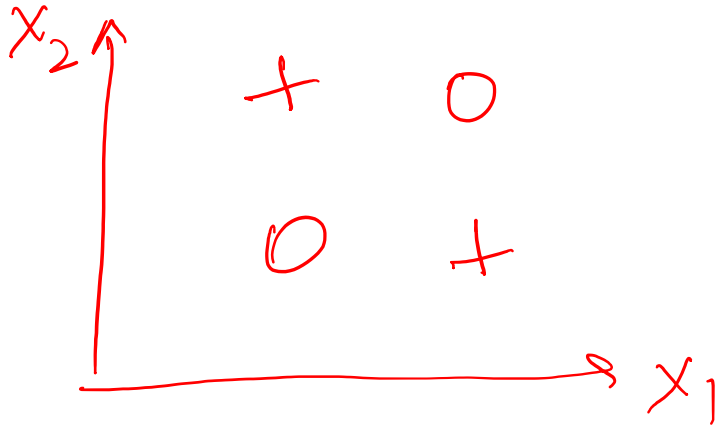
Practice problems (classification)

- To predict the chance that Steelers football team will win the Super Bowl Championship next year, you should prefer to use logistic regression instead of ~~decision trees~~. k-nearest neighbor

$P(Y=1|X)$

- Which of the following classifiers can perfectly classy the following dataset

  - Naïve Bayes No.
  - Logistic regression No.
  - Neural network ✓



$\sum_i w_i x_i \geq 0$

$\sum w_i x_i \geq 1$

$e^{\sum w_i x_i}$

$1+e^{\sum w_i x_i}$

Practice problems (regression)

- Suppose you wish to predict age of a person from his/her brain scan using regression, but you only have 10 subjects and each subject is represented by the brain activity at 20,000 regions in the brain. You would prefer to use least squares regression instead of ridge regression.

$n = 10$ $\quad dim = 20,000$

- When doing kernel regression on a memory-constrained device, you should prefer to use a box kernel instead of a Gaussian kernel.