Course Overview and Exam Review

Mathematics for Machine Learning

Lecturer: Matthew Wicker

Logistics: Exam Review

Lecture notes errors will be corrected in blue

Practice exam and equation sheet now on Scientia

Review lecture today + problem review sessions

Prize raffle postponed until next Friday: please continue to send notes errors!

I will be much more active on Ed-Stem for the next two weeks

A few notes on materials

I still owe you lecture notes for Lecture 10

Answers for Lecture 9

Answers to 3 Questions from Lecture 1

Answers to the practice exam will be posted on Monday afternoon

I have not completed my correction of all the notes, but a few have been updated. Edstem post when I am done!

Lecture 1 Material to Review

- Independent and identically distributed random variables
- Summation notation/Linear algebra review
- ML problem settings: Regression, classification, density estimation, dimensionality reduction

Recalling our supervised learning notation

$\boxed{x \in \mathbb{R}^n}$ Feature vector/Input Space

$\boxed{y \in \mathbb{R}^m}$ Labels/Outputs/Responses/Independent Variables

$\boxed{\mathcal{D} := \{(x^{(i)}, y^{(i)}\}_{i=1}^{K}}$ Dataset. Assump:, K >> n, iid

$\boxed{\min_\theta \mathbb{E}\left[\mathcal{L}(f^\theta(x), y)\right]}$ Objective

Unsupervised Learning

$\boxed{x \in \mathbb{R}^n}$ Feature vector/Input Space

$\boxed{\{x_1, x_2, \ldots, x_N\}}$ Dataset. Assump:, K >> n, iid

Here we have no response variables/labels. This is the counter to supervised learning and like the many supervised learning settings we have seen, unsupervised learning has a myriad of tasks under its umbrella

Lecture 2 Material to Review

- Vector calculus: Computing derivatives, gradients, Hessians
- Ordinary least squares estimation
  - Write loss in matrix form
  - Take the derivative
  - Set equal to zero and solve
- Basis expansion

Eigen Decomposition

$$A \in \mathbb{R}^{n \times n}$$

$$A = Q \Lambda Q^{-1}$$

Columns are the eigenvectors of A

Diagonal matrix with each entry corresponding to eigenvalues

Deriving the optimal parameter value

$$\nabla_\theta ||\theta^\top \mathbf{X} - \mathbf{y}||_2^2 = -2\mathbf{X}^\top (y - \theta^\top \mathbf{X})$$

$$-2\mathbf{X}^\top (y - \theta^\top \mathbf{X}) = 0$$

$$\mathbf{X}^\top y - \mathbf{X}^\top \mathbf{X} \theta^\top = 0$$

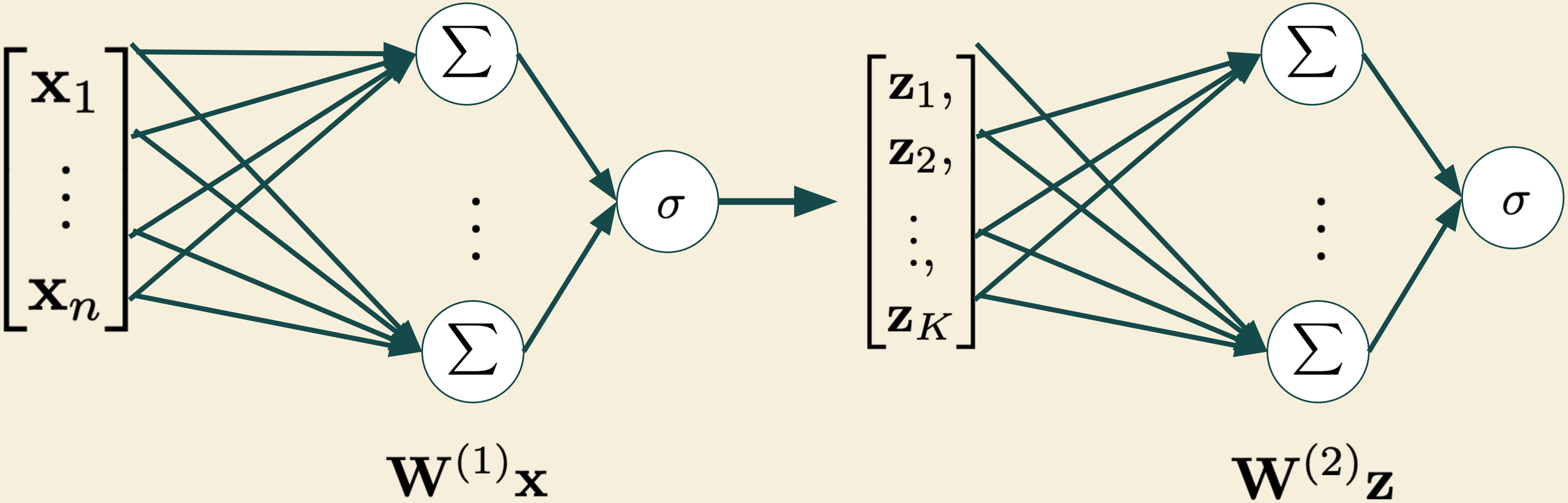
$$\mathbf{X}^\top y = \mathbf{X}^\top \mathbf{X} \theta^\top$$

$$(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top y = \theta^\top$$

Lecture 3 Material to Review

- Automatic differentiation
- Forward-mode
- Reverse-mode
- Constructing computational graphs
- Computational complexity of forward and reverse mode
- Fully connected neural networks

Formulating the MLP



Forward Mode Auto. Diff.

$$
\begin{aligned}
\mathbf{J}\mathbf{x} &= \mathbf{J}^{(L)}\mathbf{J}^{(L-1)}\ldots\mathbf{J}^{(2)}\left(\mathbf{J}^{(1)}\mathbf{x}\right)\\
&= \mathbf{J}^{(L)}\mathbf{J}^{(L-1)}\ldots\mathbf{J}^{(3)}\left(\mathbf{J}^{(2)}\mathbf{x}^{(1)}\right)\\
&= \mathbf{J}^{(L)}\mathbf{J}^{(L-1)}\ldots\mathbf{J}^{(4)}\left(\mathbf{J}^{(3)}\mathbf{x}^{(2)}\right)\\
&\ldots\\
&= \mathbf{J}^{L}\mathbf{x}^{(L-1)}
\end{aligned}
$$

Continue until we have just the final Jacobian

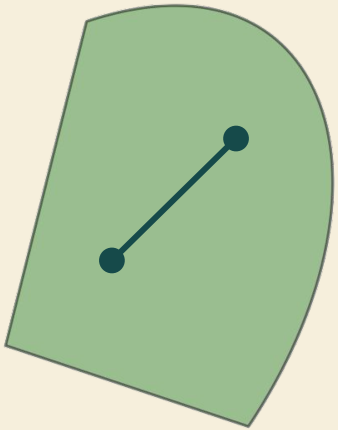
Lecture 4 Material to Review

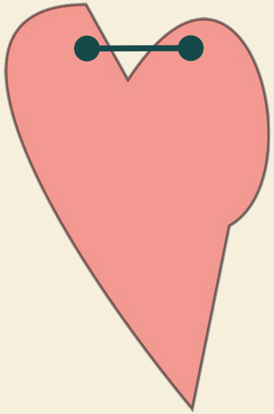
- Definition of convergence
- Definition of convexity
- Gradient descent algorithm
- Complexity of gradient descent algorithm
- Convergence analysis of gradient descent
- *Lipschitz continuity is not examinable*

Convex Sets

$C \subseteq \mathbb{R}^n$ is convex if $\forall x, y \in C$ and $\forall t \in 0 \leq t \leq 1$

$$tx + (1-t)y \in C$$

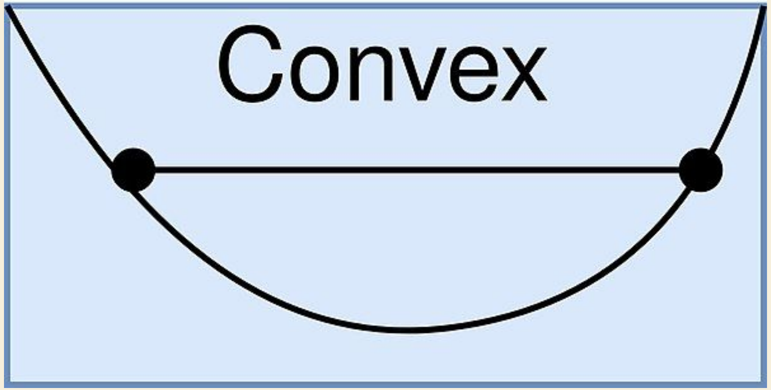


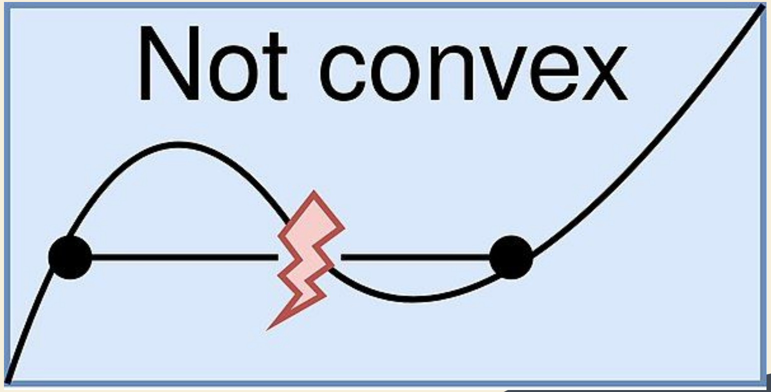


Convex Functions

$$\forall x, y \in \mathrm{dom}(f), \forall t \in [0, 1]$$

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$





Lecture 5 Material to Review

- Maximum likelihood estimation
  - For density estimation
  - For conditional density estimation
- Recall: use the density of the probabilistic model to derive the NLL, differentiate, solve for zero
- Maximum a posteriori estimation
- Connections to information theory is not examinable

Computing the MLE in linear regression

$$-\log\left(p(\mathcal{D}|\theta)\right) = \boxed{\frac{N}{2}\log(2\pi\sigma^2)} + \boxed{\frac{1}{2\sigma^2}}\boxed{(\mathbf{X}\theta - \mathbf{y})^\top(\mathbf{X}\theta - \mathbf{y})}$$

We arrive at just the OLS estimator using the NLL in MLE:

$$(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top y$$

Maximum a posteriori

$$\frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left(-\frac{(\mathbf{y}-\mathbf{X}\theta)^\top(\mathbf{y}-\mathbf{X}\theta)}{2\sigma^2}\right)\frac{1}{(2\pi\tau^2)^{d/2}}\exp\left(-\frac{\theta^\top\theta}{2\tau^2}\right)$$

Now to find the argmax parameter for this model we need to follow our steps from our MLE exposition:

1. Go from this likelihood to the negative log likelihood (NLL)
2. Set equal to zero and solve for theta

$$\theta^{\mathrm{MAP}} = (\mathbf{X}^\top\mathbf{X} + \frac{\sigma^2}{\tau^2}\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$$

Lecture 6&7 Material to Review

- Basic measure theory definitions of probability terms (PDF, CDF)
- Joint probability, marginal probability, conditional probability
- Conditional independence
- Law of total expectation, law of total variance
- Change of variables
- LOTUS

Conditional vs. Marginal

$P(X=x)$

$p(x, y)$	Type 1	Type 2	Total
Malignant	4	8	12
Benign	7	9	16

0.42
0.58

Marginal distribution

0.36
0.64

Conditional distribution

Change of variables

$$Y = g(X)$$

A brand new random variable!

The distribution of this random variable is:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right|$$

Law of the unconscious statistician

$$\mathbb{E}_X[f(X)] = \sum_x \left(\sum_{z:T(z)=x} p_Z(Z=z) \right) f(x)$$

$$= \sum_z p_Z(Z=z) f(T(z)) = \mathbb{E}_Z[f(T(Z))]$$

$$\boxed{\mathbb{E}_X[f(X)] = \mathbb{E}_Z[f(T(Z))]}$$

Lecture 8 Material to Review

- Bayes theorem
  - Posterior inference for density estimation
  - Posterior inference for conditional density estimation
- Interpretation and role of terms in Bayes theorem:
  - Prior
  - Likelihood
  - Marginal likelihood/evidence
- Conjugacy of prior distributions — Beta bernoulli & Gaussian only

Everyday probabilistic reasoning

Your young cousin is enthusiastic about birds and wants you to help them identify a cool bird they have just seen. They describe it as being white with an orange/red-ish beak

In reality it is much closer to being flipped!

$P(\text{Gull} \mid \text{Obs.})$	$P(\text{Pigeon} \mid \text{Obs.})$
~~0.7~~ 0.3	~~0.3~~ 0.7

Method: Equating coefficients

$$p(s|v) \propto p(v|s)p(s)$$

$$= \mathcal{N}(v; s, \sigma^2)\mathcal{N}(s; 0, 1)$$

$$= \exp\left(-\frac{v^2}{2\sigma^2} + \frac{sv}{\sigma^2} - \frac{s^2}{2\sigma^2} - \frac{s^2}{2}\right)$$

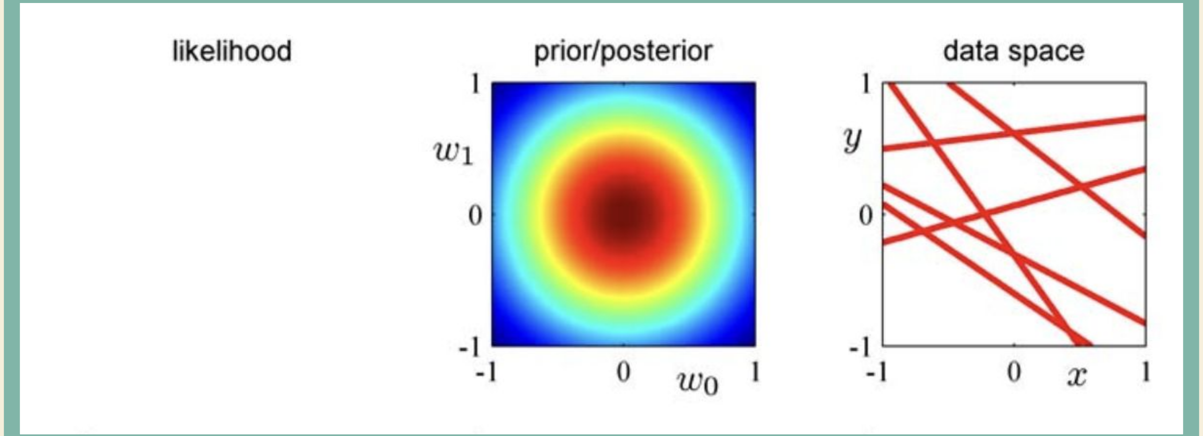
$$\propto \exp\left(-\frac{1+\sigma^2}{2\sigma^2}s^2 + \frac{v}{\sigma^2}s\right)$$

Key idea:
Conjugacy!

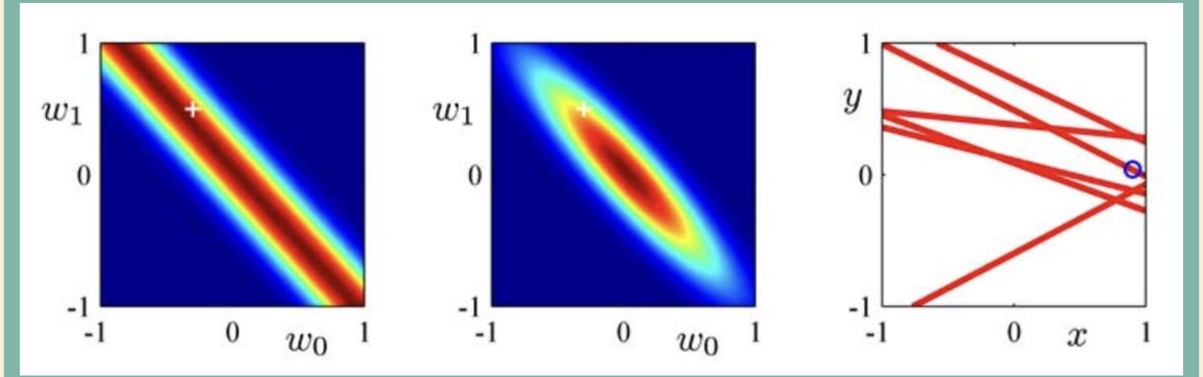
Lecture 9 Material to Review

- All of lecture 8 material but applied to linear regression
- Using the method of equating coefficients
- Using the method of joint Gaussians
- Deriving the posterior predictive distribution
- Woodbury Identity

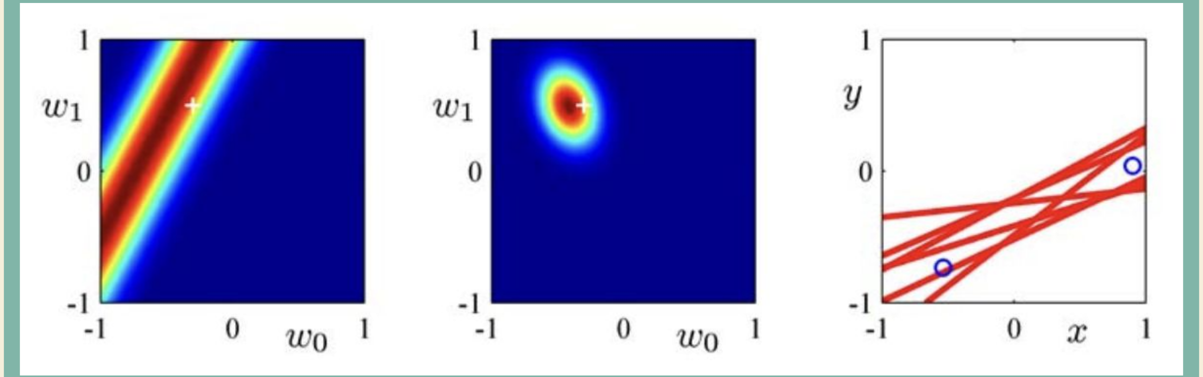
Epistemic uncertainty in ML is easier to think about in "function" space rather than strictly in "parameter" space.



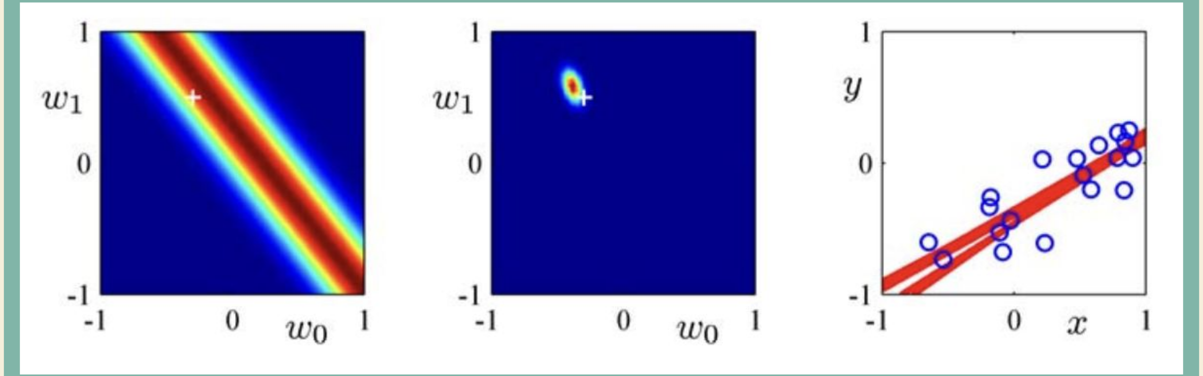
Epistemic uncertainty in ML is easier to think about in "function" space rather than strictly in "parameter" space.



Epistemic uncertainty in ML is easier to think about in "function" space rather than strictly in "parameter" space.



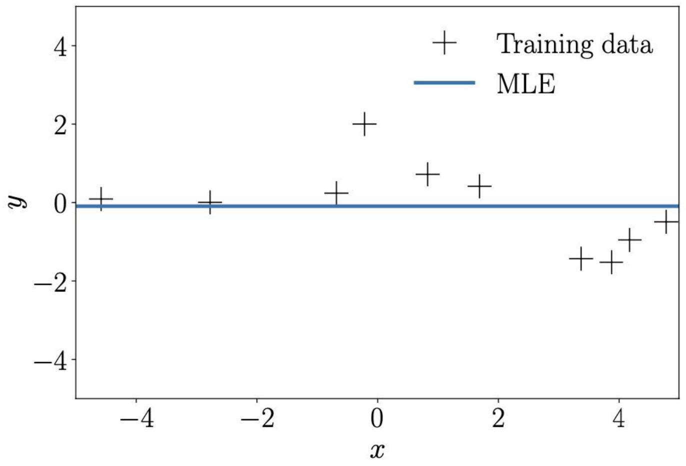
Epistemic uncertainty in ML is easier to think about in "function" space rather than strictly in "parameter" space.



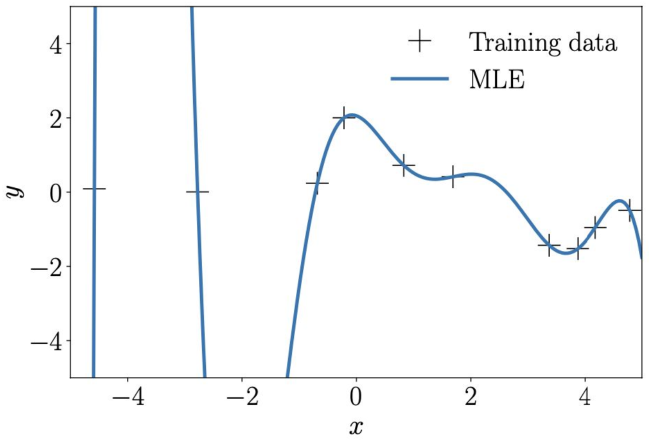
Lecture 10 Material to Review

- Measuring generalization with a test-set
- Markov + Chebyshev's inequality
- Weak law of large numbers
- Identifying overfitting
- Universal function approximation is not examinable

What is happening to our loss?



$\boldsymbol{\phi}(x) = [1]^\top$



$\boldsymbol{\phi}(x) = [1\ x\ x^2\ x^3, \ldots]^\top$

What is the mathematical object?

Training Data	Test data

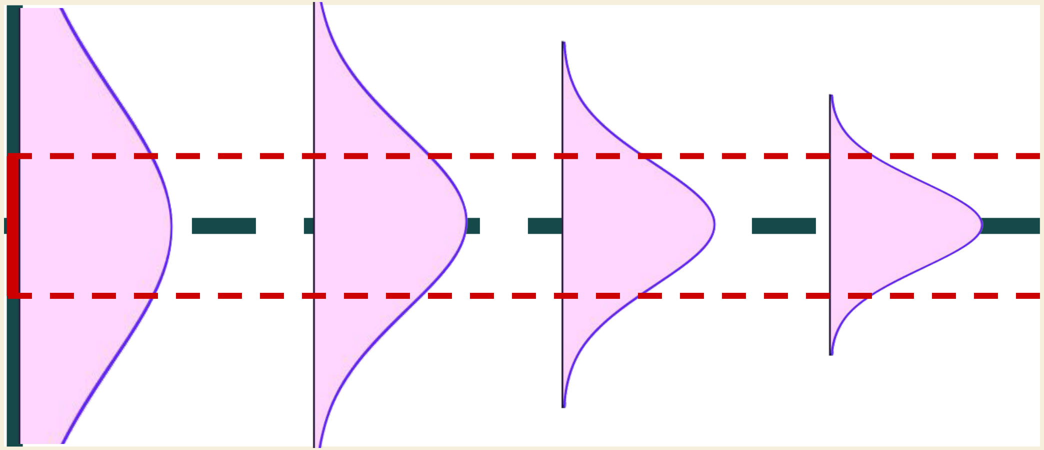
$$\{(X_1 = x_1, Y_1 = y_1), (X_2 = x_2, Y_2 = y_2), \ldots, (X_N = x_N, Y_N = y_N)\}$$

$$\{Z_1 = \ell(f^\theta, X_1 = x_1, Y_1 = y_1), \ldots, Z_N = \ell(f^\theta, X_N = x_N, Y_N = y_N)\}$$

What assumptions do we need to make here?

Convergence of Sequence of R.V.:

$$\forall \epsilon > 0, \lim_{n \to \infty} P(|Z_n - a| \geq \epsilon) = 0$$



$$\forall \epsilon > 0, \forall \epsilon' > 0, \exists M \text{ s.t. } \forall M' > M, P(|Z_n - a| \geq \epsilon) \leq \epsilon'$$

We call this the weak law of large numbers

$$\frac{\sigma^2}{a^2} \geq P(|z - \mu| \geq a)$$

Note this isn't exactly Chebyshev, but it the step just before (so equivalent)

$$P(|\hat{Z} - \mu| \geq \epsilon) \leq \frac{\mathbb{V}[\hat{Z}]}{\epsilon^2} = \frac{\sigma^2}{N\epsilon^2}$$

Lecture 11 Material to Review

- Hoeffding's inequality
- Regularization
  - Implicit regularization
  - Explicit regularization
- Cross validation
  - K-fold cross-validation
  - LOOCV
- Complexity and trade-offs of cross-validation
- PAC generalization bound is not examinable

Explicit Regularization

$$\mathcal{L}_{\text{reg.}}(\theta) := \mathcal{L}(\theta) + \alpha C(\theta)$$

Explicit regularization is characterized by the addition of terms to our loss function that encode constraints over the parameters or outputs of our ML model. Weight decay:

$$\mathcal{L}_{\text{ridge}}(\theta) := \mathcal{L}(\theta) + \alpha||\theta||_2$$

Implicit Regularization

- reasoning about this not in exam

ML Model (e.g., neural network)

Decisions we need to make

Loss function

Dropout

Optimizer

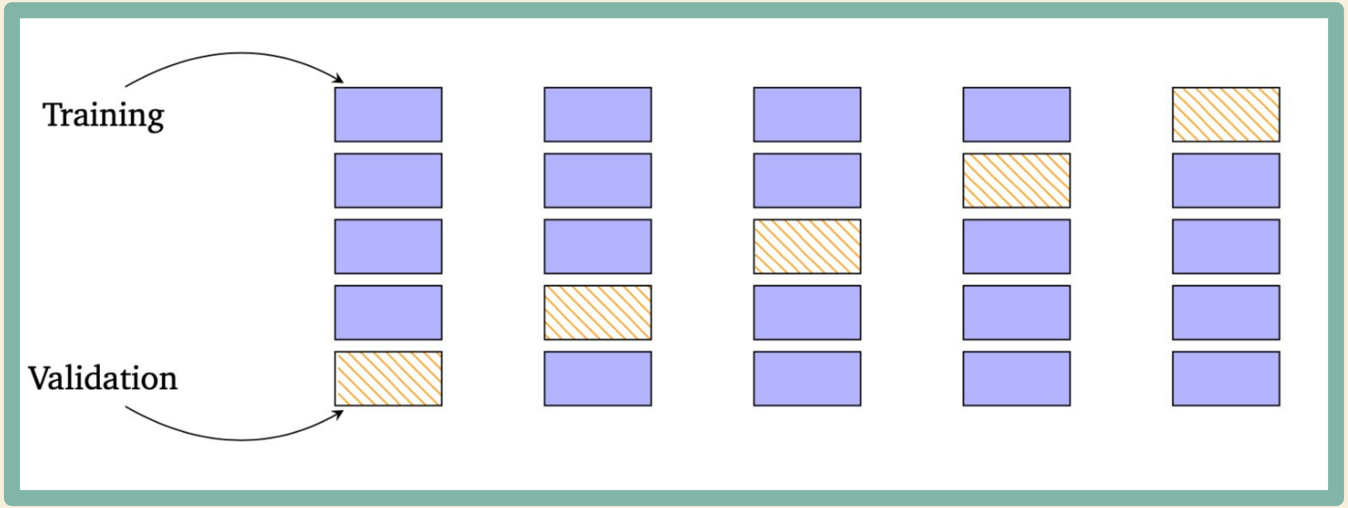
Number of layers

Epochs (Early Stopping)

Width of Layers

(K fold) Cross-validation

The key idea behind cross-validation is to split our training data into K mutually exclusive subsets each of which will be used as the validation in one of K seperate algorithm runs



Lecture 12 Material to Review

- Definition of an estimator
- Bias of an estimator
- Variance of an estimator
- Bias-variance trade-off
- Bias-variance decomposition in our studied models
    - Using this to motivate regularization

We talk of estimators, let's define them

Definition 2.1. Statistic A statistic S is a random variable that is a function of some data $\mathcal{D}$, $S = g(\mathcal{D})$ where the data $\mathcal{D}$ is a collection of random variables.

An estimator is a statistic that aims at approximating a parameter/property of a distribution.

$$\hat{Z} = \frac{Z_1 + Z_2 + \ldots + Z_N}{N}$$

We talk of estimators, let's define them

An estimator is a statistic that aims at approximating a parameter/property of a distribution.

$$\mathrm{Bias}(\hat{Z}_n) = \mathbb{E}[\hat{Z}_n - Z] = 0$$

Unbiased estimator

Variance of an estimator

Definition 2.1. Statistic A statistic S is a random variable that is a function of some data $\mathcal{D}$, $S = g(\mathcal{D})$ where the data $\mathcal{D}$ is a collection of random variables.

An estimator is a statistic that aims at approximating a parameter/property of a distribution.

$$\mathrm{Var}(\hat{S}) = \mathbb{E}\left[(\hat{S} - \mathbb{E}(\hat{S}))^2\right]$$

Lecture 13 Material to Review

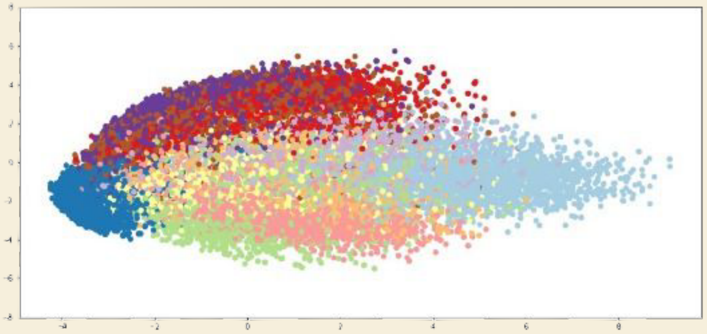
- Dimensionality reduction (the problem setting)
- PCA - Maximizing the variance formulation
- PCA - Minimum reconstruction error formulation
- Being able to apply all concepts from the course to a PCA set up.

Non-linear dimensionality reduction

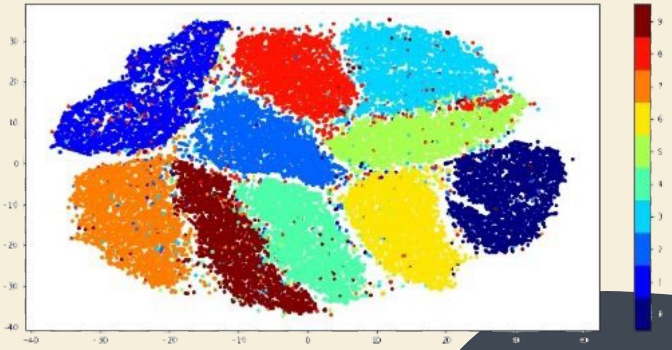
Algorithm 1 PCA
Input: $\boldsymbol{X}$ - Feature/Design matrix, K - Number of components

$\hat{S}_n \leftarrow \sum_{i=1}^{n} \boldsymbol{x}^{(i)} \boldsymbol{x}^{(i)\top}$
Compute eigendecomposition $S = P \Lambda P^\top$
Ensure $\Lambda = \mathrm{diag}(\lambda)$ with $\lambda_i \geq \lambda_j \forall i < j$
$B = P_{[:,:k]}$
return $\{B\boldsymbol{x}^{(i)}\}_{i=1}^{n}$

MNIST - PCA



MNIST - TSNE



Next lecture: None

Next lecture: None
(But some research on the board now)