Machine Learning

Lec 2: Theory Foundation

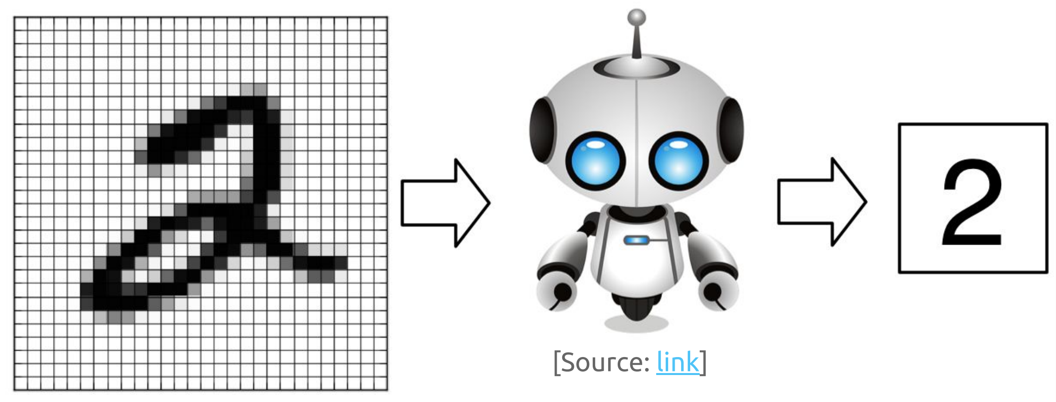
Prof. Da-Cheng Juan

Agenda

- Latex/Project setup
- Recap: Supervised Learning
- Dataset
- Probability
- MLE
- MAP
- Model complexity and overfitting

Magic Behind the Scene

- Supervised Learning
  - Mostly, also unsupervised learning. Wait, where is deep learning?

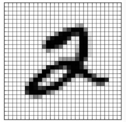
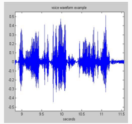


Features
(referred as "**x**")

Function
(Not a robot, yet)

Labels
(referred as "y")

Just a Function

- Handwriting recognition: *f* () = “2”
- Speech recognition: *f* () = “吃飽了沒?”
- Playing Go: *f* () = “4-5” (next move)
- How should we pick “*f*”?
    - **Deep Neural Nets** (sounds easy!)

Supervised Learning

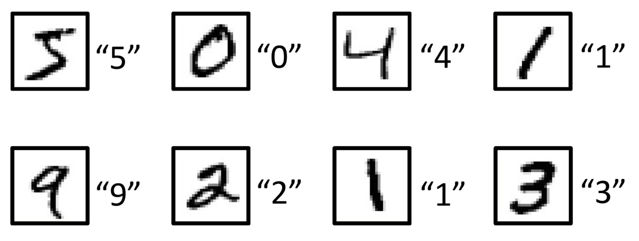
What you should know:

- Well posed function approximation problems:
  - Instance space, X
  - Sample of labeled training data { $<x^{(i)}, y^{(i)}>$}
  - Hypothesis space, H = { f: X→Y }

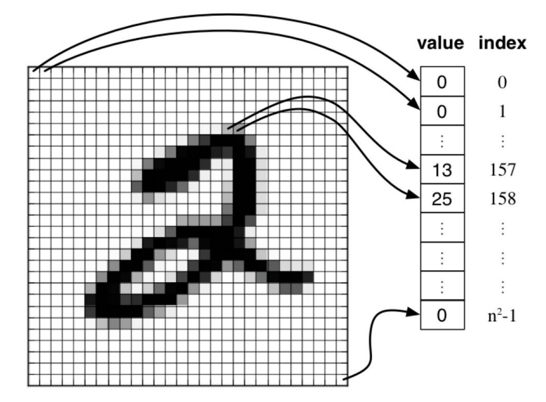
- Learning is a search/optimization problem over H
  - Various objective functions
    - minimize training error (0-1 loss)
    - among hypotheses that minimize training error, select smallest (?)
  - But inductive learning without some bias is futile !

Data Representation

- Dataset usually contains lots of samples
  - Otherwise hard to learn
- Each sample contains:
  - Features: $\mathbf{x}$ (x_1, …, x_k)
  - Label: y



[Source: Hung-Yi Lee, http://www.slideshare.net/tw_dsconf/ss-62245351]



Label: “2”

Categorical v.s. Ordinal Features

- Categorical features (or labels)
  - Contain two or more categories, and no intrinsic ordering exists in them.
  - For example: blood types (“A”, “B”, “AB”, and “O”)
  - Require additional encoding, usually one-hot encoding (illustration).
- Ordinal Features (or labels)
  - Ordinal features refer to quantities that have a natural ordering.
  - For example: a temperature of 35°C is larger (or hotter) than 33°C.

Pre-process Dataset: Min-Max Normalization

- Benefit of pre-processing:
  - Learn faster. Why?
  - Prevent numerical error during training.

- Min-Max Normalization:
  - Re-scaling the range of a vector to make all the elements lie between 0 and 1, a.k.a., “Min-max normalization.”
  - $$x' := \frac{x - x_{min}}{x_{max} - x_{min}}$$

Pre-process Dataset: Standardization

- Subtract the mean and divide by the standard deviation.

$$x_i' := \frac{(x_i - \mu)}{\sigma}$$

- When to use which?
  - Min-max normalization or standardization?

Probability Overview

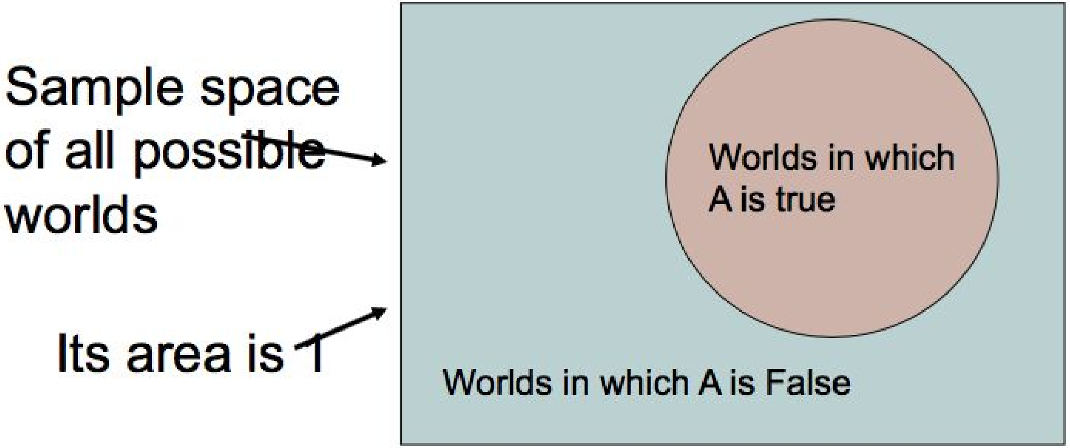
- Events
  - discrete random variables, continuous random variables, compound events
- Axioms of probability
  - What defines a reasonable theory of uncertainty
- Independent events
- Conditional probabilities
- Bayes rule and beliefs
- Joint probability distribution
- Expectations
- Independence, Conditional independence

Formulation

More formally, we have

- a sample space S (e.g., set of students in our class)
  - aka the set of possible worlds
- a random variable is a function defined over the sample space
  - Gender: S → { m, f }
  - Height: S → Reals
- an event is a subset of S
  - e.g., the subset of S for which Gender=f
  - e.g., the subset of S for which (Gender=m) AND (eyeColor=blue)
- we're often interested in probabilities of specific events
- and of specific events conditioned on other specific events

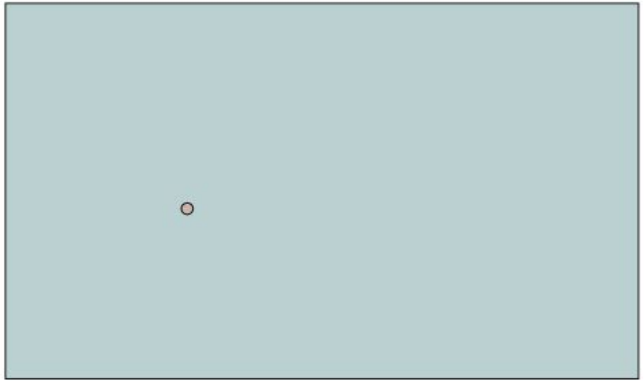
Visualization



P(A) = Area of reddish oval

Axiom

- 0 <= P(A) <= 1
- P(True) = 1
- P(False) = 0
- P(A or B) = P(A) + P(B) - P(A and B)

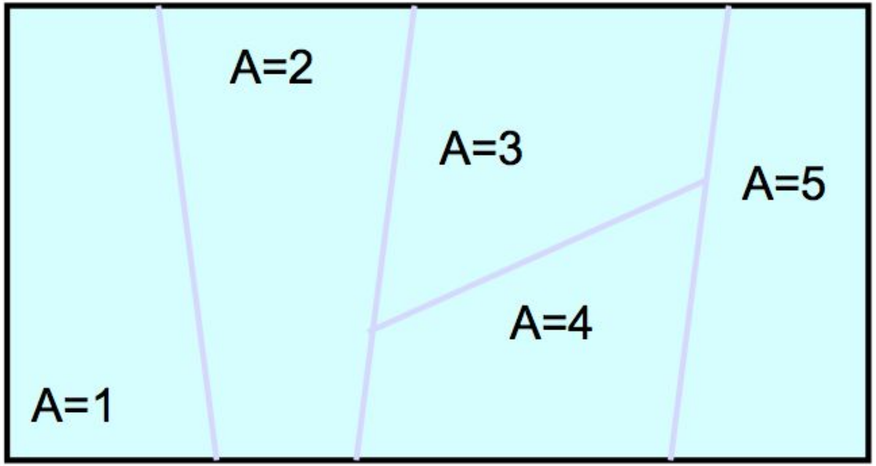


The area of A can't get any smaller than 0

And a zero area would mean no world could ever have A true

Elementary Probability

$$\sum_{j=1}^{k} P(A = v_j) = 1$$



Conditional Probability

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

Corollary: The Chain Rule

$$P(A \wedge B) = P(A|B)\, P(B)$$

Your First Consulting Job:

- A billionaire from the suburbs of Seattle asks you a question:
  - He says: I have thumbtack, if I flip it, what's the probability it will fall with the nail up?
  - You say: Please flip it a few times:

  - You say: The probability is:
  - **He says: Why???**
  - You say: Because…

Binomial Distribution

- P(Heads) = θ, P(Tails) = 1-θ

D: 1 1 0 1 0

X_1 X_2 X_3 X_4 X_5

- Flips are i.i.d.:
  - Independent events
  - Identically distributed according to Binomial distribution
- Sequence D of α_H Heads and α_T Tails

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

Maximum Likelihood Estimate (MLE)

- **Data:** Observed set D of α_H Heads and α_T Tails
- **Hypothesis:** Binomial distribution
- Learning θ is an optimization problem
  - What's the objective function?

- MLE: Choose θ that maximizes the probability of observed data:

$$\begin{aligned} \hat{\theta} &= \arg\max_{\theta} \; P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \; \ln P(\mathcal{D} \mid \theta) \end{aligned}$$

MLE for \theta

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} \; \ln P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \; \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}\end{aligned}$$

- Set derivative to zero: $\boxed{\frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) = 0}$

MLE for \theta (cont'd)

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} \; \ln P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \; \ln \theta^{\alpha_H} (1-\theta)^{\alpha_T}\end{aligned}$$

Maximum Likelihood Estimate: Procedure

- Write down probability for one observation
- Write down likelihood
- Calculate log-likelihood
- Take partial derivative and set to 0
    - How about not the closed form?
    - Use gradient descent instead :)

- If you forget, remember “coin flips” as example.

Bayes Rule

- Use Bayes rule:

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}$$

- Or equivalently:

$$P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$$

Maximum A Posteriori (MAP)

Maximum a Posteriori (MAP) estimate: choose θ that is most probable given prior probability and the data

$$\hat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D})$$

$$= \arg\max_{\theta} = \frac{P(\mathcal{D} \mid \theta) P(\theta)}{P(\mathcal{D})}$$

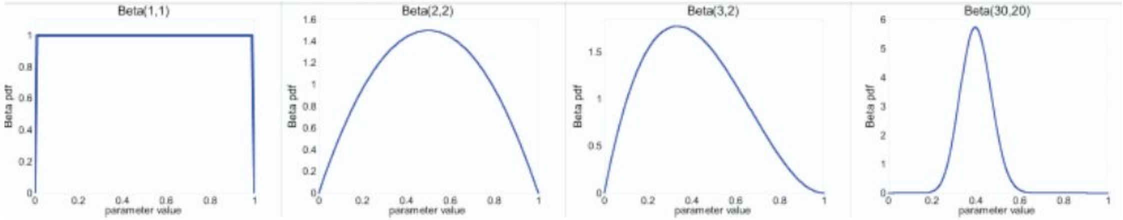
Using Prior

Beta prior distribution – P(θ)

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

Mean:

Mode:

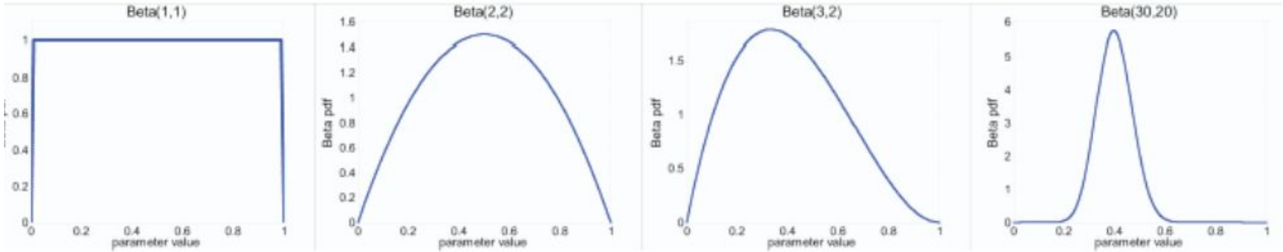


- Likelihood function: $P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$
- Posterior: $P(\theta \mid \mathcal{D}) \;\propto\; P(\mathcal{D} \mid \theta) P(\theta)$

Posterior Distribution

- Prior: $Beta(\beta_H, \beta_T)$
- Data: α_H heads and α_T tails
- Posterior distribution:

$$P(\theta \mid \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$



Putting Together: Maximum A Posteriori (MAP)

$$P(\theta \mid \mathcal{D}) = \frac{\theta^{\beta_H+\alpha_H-1}(1-\theta)^{\beta_T+\alpha_T-1}}{B(\beta_H+\alpha_H, \beta_T+\alpha_T)} \sim Beta(\beta_H+\alpha_H, \beta_T+\alpha_T)$$

- MAP: use most likely parameter:

$$\widehat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D}) =$$

- Beta prior equivalent to extra thumbtack flips
- As $N \rightarrow \infty$, prior is "forgotten"
- **But, for small sample size, prior is important!**

Parameter Estimate

- Maximum Likelihood Estimate (MLE): choose θ that maximizes probability of observed data $\mathcal{D}$

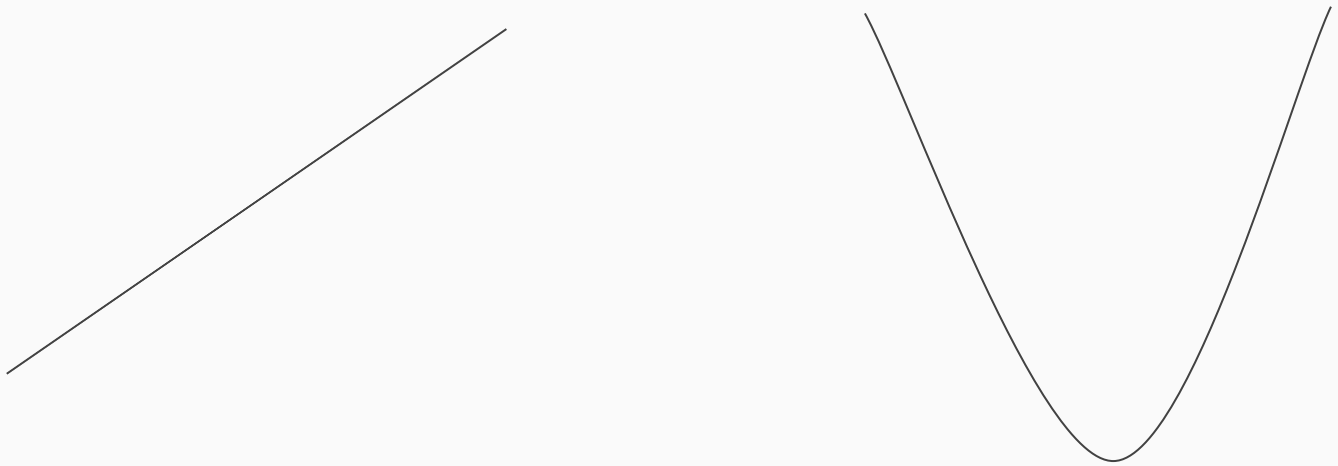
$$\widehat{\theta} = \arg\max_{\theta} P(\mathcal{D} \mid \theta)$$

- Maximum a Posteriori (MAP) estimate: choose θ that is most probable given prior probability and the data

$$\widehat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D})$$

$$= \arg\max_{\theta} = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}$$

Model Complexity: Linear Regression



Model Complexity: Linear Regression

Overfitting

Overfitting

Consider error of hypothesis h over

- training data: $error_{train}(h)$
- entire distribution $\mathcal{D}$ of data: $error_{\mathcal{D}}(h)$

Hypothesis $h \in H$ **overfits** training data if there is an alternative hypothesis $h' \in H$ such that

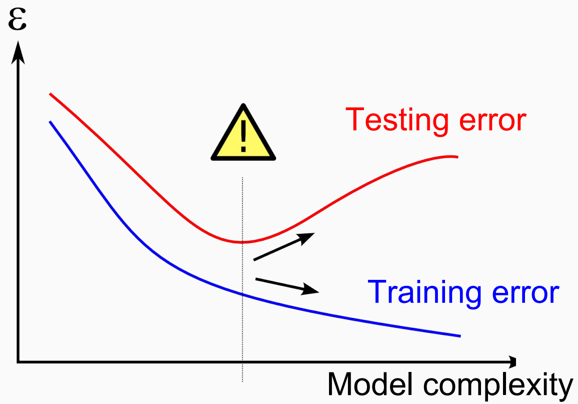
$$error_{train}(h) < error_{train}(h')$$

and

$$error_{\mathcal{D}}(h) > error_{\mathcal{D}}(h')$$

Overfitting? Training and Testing

- What is overfitting?
  - Model memorizes all the training data (especially noises), and
  - Doesn’t perform well on new incoming samples (testing data).
  - How about underfitting?
    - Very rare for DNN
  - “Kaggle example”
- How to avoid?
  - Dropout
  - Regularization
  - And other techniques, stay tuned.



[Source: https://commons.wikimedia.org/wiki/File:Overfitting.png]

Q & A