Aside on proofs:

assumptions \implies \text{ logical steps} \implies \text{ conclusion}

Example of them: \( \forall x \forall y \implies f(x, y) = p(x) \quad A \forall y \)
\( \forall x \forall y \) (assumption)

by definition \( \implies \) (\( \forall x \) \( p(x, y) = p(x) f(y) \quad A \forall y \))

by definition

\( \forall y \text{ true if } y \neq 0 \)

\( f(y) \)

\( \frac{p(x, y)}{p(y)} = \frac{p(x)}{p(y)} \quad \frac{p(x) f(y)}{p(y)} = p(x) \quad \text{if only if (equivalent)} \)

(proof by contradiction: use fact that \((A \implies B) \iff (\neg B \implies \neg A)\))

today: statistical decision theory

(frequentist) statistical decision theory

unknown distribution which models the "world"

formal setup:

- random observation \( D \sim p \) (perhaps \( f_\theta \))
- action space \( \mathcal{A} \)
- loss: \( L(D, \alpha) = \text{loss of action action} \alpha \) \( \mathbb{R} \) describes the
If you have parametric model in mind, often write $L(\theta, \alpha)$ where $\theta \in \Theta$ is set. $\hat{\alpha} = \arg \min L(\theta, \hat{\alpha})$

$S: \Theta \rightarrow A$ "decision rule"

- $\Theta$ set of parameters for $\theta_0$

Examples:
- $\hat{\alpha} = \hat{\theta}$ for a parametric family $\mathcal{P}_\theta$
  - $S$ is then parameter estimator from data
  - Typical loss $L(\theta, \alpha) = \| \theta - \alpha \|^2$ "squared loss"

More specifically:
- $\mathcal{D} = (x_1, \ldots, x_n)$ where $x_i \in \mathcal{P}_\theta$ ($\theta$ is unknown)

\[
\hat{S}(\mathcal{D}) = \hat{\theta} \quad L(\theta, \hat{\theta}(\mathcal{D})) = \| \theta - \hat{\theta} \|^2 \]

Suppose $\mathcal{P}_\theta$ is Gaussian family $\mathcal{N}(\mu, \Sigma) \Rightarrow \mu \in \mathbb{R}^2$

- $\Theta = \{ \mu \}$ (set of possible means)

\[
\hat{S}(\mathcal{D}) = \frac{1}{n} \sum_{i=1}^{n} x_i
\]

b) $\mathcal{A} = \{0, 1\}$; this is basically hypothesis testing

Here $S$ describes a statistical test

c) Machine learning: learning a prediction rule

Here $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ $x_i \in \mathcal{X}$ (input space) $y_i \in \mathcal{Y}$ (output space)

e.g. $y = \mathcal{Y}_{[0, 1]}$ for binary classification

If $S_0$ gives joint on $(x_i, y_i)$, then $\mathcal{D} \in \mathcal{D}$ where $\mathcal{D} = \mathbb{R}^2 \times \mathbb{R}^d \times \mathbb{R}^d \times \cdots \mathbb{R}^d$
\[ A = \mathcal{Y}^X \text{ (set of functions from } X \rightarrow \mathcal{Y}) \text{ (red model)} \]

in machine learning, \[ L(P, S) = \mathbb{E}_p \left[ L(Y, S(x)) \right] \]

"generalization error"

- decision rule \[ S = S(D) \]
  - prediction function
  - training dataset

"learning algorithm"

procedure analysis:

given this framework, how do we compare procedures? eg \( S_1 \) vs \( S_2 \)

\[ (\text{frequentist risk}) \quad R(p, S) = \mathbb{E}_p \left[ L(p, S(D)) \right] \]

"risk profiles"

\[ S_1 \quad S_2 \]

"how well \( S \) does in average"

how to go from risk profile to a scalar number

\[ R \]

\[ \text{``minimum''} \]

\[ \Rightarrow \text{look at max } R(p, S) \]

alternative in ML theory is PAC theory - blockade at bad bounds
Bayesian decision theory

\[ p(x) \]

... "prior" over \( p \)

\[ \sum_{x} p(x, \theta) \pi(c) \, dx \]

\[ \mathbb{P}[L(p, S_0)] \leq \text{threshold} \]

"with high prob" statement

\[ \text{Contour:} \]

\[ L(p, S_0) \]

when \( D \) reduces \( p \)

\[ R(p, \theta) \]

Bayesian decision theory

\[ \mathbb{P}_B(a | D) = \int L(g, a) \, \pi(g | D) \, dg \]

posterior \( \pi(g | D) \)

Bayesian posterior risk

Bayesian optimal action

\[ \text{Bayesian}(D) = \text{argmax } \mathbb{P}_B(a | D) \]

... "estimation"

Example: \( g, a = 10 \)

\[ L(g, a) = 1/10 - a^2 \]

Then (exercise)

\[ \text{Bayesian}(D) = E[B | D] \]

(posterior mean)

\[ L(\theta, S_0) \]

Teaching Page 4
Examples of estimators: $E_S = \hat{p}(\theta)$

- MLE

  - another: MAP, given a prior $p(\theta)$
    
    $$\text{then pick } \hat{\theta} = \arg\max_{\theta \in \Theta} p(\theta | D)$$
    
    $$\propto p(D | \theta) p(\theta)$$

- method of moments

  - idea: find an injective mapping from $\Theta$ to moments
    
    and say estimate on "possible moments"
    
    and then invert it from empirical moments

  - example: for Gaussian $X \sim N(\mu, \sigma^2)$
    
    $$\mathbb{E}[X] = \mu$$
\[ \mu(x) = \sigma^2 + \mu^2 \]

\[ \mathbb{E}[\mu(x)] = f(x, \sigma^2) \]

\[ \mathbb{E}[\sigma^2] = \mathbb{E}[\mathbb{E}[\sigma^2 | x]] \]

(here, this same as MLE (property of exponential family))

- This is useful for latent variable models (e.g., mixture of Gaussians)
  
  (special methods)
  
  e.g.,

  - in the context of prediction \( \hat{f} = \mathbb{E}[f(x) | s] \)
    
    \( x \in \text{input space} \)
    
    \( s \in \text{output space} \)

  example of \( S: D \rightarrow R \)

  is using empirical risk minimization (ERM)

  (empirical risk) i.e. generalization over

  model: \[ L(p, S) = \mathbb{E}_{(x, y) \sim p} \mathbb{E}[L(y, f(x))] \]

  replace \[ \mathbb{E} \mathbb{E}[L(y, f(x))] = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) \]

  \[ \hat{f}_{\text{ERM}} = \text{argmin}_{f \in \mathcal{F}} \mathbb{E} \mathbb{E}[L(y, f(x))] \]

  \[ f \in \mathcal{F} \text{ hypothesis class} \]