Machine Learning (CS 181):
Style Guide

Basic Notation

$\mathbf{A} \in \mathbb{R}^{m \times n}$	matrices are bold caps
$\mathbf{a} \in \mathbb{R}^{m(\times 1)}$	vectors are bold lower, always column
$\mathbf{a}^\top$	transpose
$a \in \mathbb{R}$	scalars are lower case, non bold
$\mathcal{A}$	sets are script case
$\{\mathbf{a}_i\}_1^n$	a sequence of $\mathbf{a}_1 \ldots \mathbf{a}_n$

- We distinguish between $\mathbf{a}_i$ and b_i. The first is the $i'th$ vector of a sequence, the second is the i'th scalar in $\mathbf{b}$.

Supervised Learning

$\mathbf{X} \in \mathbb{R}^{n \times m}$	n training instances with m features
$X_{i,j}$	the j'th feature of example i
$\mathbf{y} \in \mathbb{R}^n$	the n target instances (regression)
$\mathbf{y} \in \{-1, 1\}^n$	the n target instances (classification)

- When describing data as a sequence we use
  $D = (\mathbf{x}_1, y_1) \ldots (\mathbf{x}_n, y_n) = \{(\mathbf{x}_i, y_i)\}_1^n$

Models

$\mathbf{w} \in \mathbb{R}^m$	linear-model parameters
$f(x; \mathbf{w})$	model parameterized by $\mathbf{w}$
$\hat{y}$	model prediction (to be distinguished from target y)
$\phi : \mathbb{R}^m \mapsto \mathbb{R}^d$	basis function (if used changes $\mathbf{w}$ dim)

Optimization

- $\mathcal{L}(\mathbf{w})$; loss function
- $\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w})$; training minimization
- $\mathcal{L}(\mathbf{w}) = -\sum_{i=1}^{n} \ln p(y_i \mid \mathbf{x}_i; \mathbf{w})$; negative log-likelihood
- $\min_x \mathbf{x} \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{b}$; constrained

Probabilistic Notation

- $p(y \,|\, x)$; discrete distribution $p(Y = y \,|\, X = x)$ (RVs implied)
- $p(y \,|\, x) \propto p(x \,|\, y) p(y)$; proportional to notation
- $\mathcal{N}(x \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})$; multivariate normal (prefer $\exp$ to e)

$$\mathcal{N}(x \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}))$$

- α, β, γ; scalar hyperparameters (explicitly conditioned $p(w|\alpha)$)
- $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$; vector hyperparameters