Classifications

(1) Generative models (e.g., LDA) [We'll learn about LDA next lecture.]
- Assume sample points come from probability distributions, different for each class.
- Guess form of distributions
- For each class C, fit distribution parameters to class C points, giving $f(X|Y = C)$
- For each C, estimate $P(Y = C)$
- Bayes' Theorem gives $P(Y|X)$
- If 0-1 loss, pick class C that maximizes $P(Y = C|X = x)$ [posterior probability]
  equivalently, maximizes $f(X = x|Y = C)\,P(Y = C)$

(2) Discriminative models (e.g., logistic regression)
[We'll learn about logistic regression in a few weeks.]
- Model $P(Y|X)$ directly

(3) Find decision boundary (e.g., SVM)
- Model $r(x)$ directly (no posterior)

Advantage of (1 & 2): $P(Y|X)$ tells you probability your guess is wrong
[This is something SVMs don't do.]
Advantage of (1): you can diagnose outliers: $f(X)$ is very small
Disadvantages of (1): often hard to estimate distributions accurately;
real distributions rarely match standard ones.

Optimization Problems

1.) Unconstrained: $\min_w / \max_w P(w)$
↳ gradient descent

2.) Constrained: Unconstrained + subject to $g(w) = 0$
↳ Lagrange

3.) Linear Program: $\min_w / \max_w c \cdot w$ subject to $Aw \leq b$
↳ Simplex

4.) Quadric Program: $\min_w / \max_w w^T Q w + c^T w$, $Ax \leq b$
↳ If Q PSD, 1 local min.

Centroid

$f(x) = (\mu_c - \mu_x) \cdot x - (\mu_c - \mu_x) \cdot \frac{\mu_c + \mu_x}{2}$

Perceptrons $(+1, -1)$

- If $x_i \cdot w \geq 0$, predict 1
  Else -1
- $L(z, y_i) = \begin{cases} 0 & \text{if } y_i z \geq 0 \\ -y_i z & \text{otherwise} \end{cases}$ * z is classifier's prediction.
  ↳ $R(w) = \frac{1}{n}\sum_{i=1}^{n} L(x_i \cdot w, y_i)$
  $= \frac{1}{n}\sum_{i} -y_i x_i \cdot w$

Find $\min_w R(w)$ w/ Gradient Descent

* $O\left(\frac{\max \|X_i\|}{\gamma^2}\right)$ iterations
where γ = max margin

SVMs

- Linear, decision boundary not too close

Hard Margin

- Constraints: $y_i(w \cdot X_i + \alpha) \geq 1$
  ↳ $dist(X_i, \text{hyperplane}) = \frac{w}{\|w\|} \cdot X_i + \frac{\alpha}{\|w\|}$

∴ margin is $\min_i \frac{1}{\|w\|} |w \cdot X_i + \alpha| \geq \frac{1}{\|w\|}$ → slab width is $\frac{2}{\|w\|}$

↳ $\min_{w,\alpha} |w|^2$ s.t. $y_i(X_i \cdot w + \alpha) \geq 1$

Soft Margin

- Allows some points to violate the margin
  ↳ Constraints: $y_i(X_i \cdot w + \alpha) \geq 1 - \xi_i$, $\xi_i \geq 0$

∴ Find $\min_{w,\alpha} |w|^2 + C\sum \xi_i$

Bayesian Risk Minimization

- $R(r) = E[L(r(X), Y)]$
  $= \sum(L(r(x), 1) P(Y=1|X=x) + L(r(x), -1) P(Y=-1|X=x)) P(X=x)$
  ↳ If L symmetric, pick class that max posterior probability

MATH Dumps

- Diagonalizable matrices → VAV^T
  ↳ if symmetric ⇒ $\vec{v}$ orthogonal ($v_i^T v_j = 0$)
- SVD → $X = USV^T$
- PSD ⇔ $x^T Ax \geq 0$
  ⇔ All $\lambda_i \geq 0\ \forall i$
  ⇔ $\exists U: A = UU^T$

- $A^2 v = A(Av) = \lambda(Av) = \lambda^2 v_0$
- $A^{-1} v = A^{-1}(\frac{1}{\lambda} Av) = \frac{1}{\lambda} v_0$
- For symmetric A^2: $VAV^T VAV^T = VA^2V^T$
- Every A^2 is PSD
- $Var(R) = Cov(R,R) = E[R^2] - E[R]^2$
- $Cov(R,S) = E[(R - E[R])(S - E[S])^T = E[RS^T] - \mu_R \mu_S^T$
- $(A - Iv) = 0$
- All Σ's are PSD.
- For Gaussian dist
  ↳ ellipsoid's radii $\propto \sqrt{\lambda}$ of Σ
- Ellipse in the form
- Gradient descent: $w_{t+1} = w_t - \varepsilon J'(w)$

GDA (QDA, LDA)

$X \sim N(\mu, \sigma^2)$: $f(x) = \frac{1}{(2\pi\sigma)^{\frac{1}{2}}} \exp\left(-\frac{\|x-\mu\|^2}{2\sigma^2}\right)$

* 1: dimension

Predict $\max_c P(X = x|Y = c)\,\pi_c$
↓ ln

QDA: $\ln((2\pi\Sigma)^{\frac{1}{2}} f_c(x) \pi_c) = -\frac{\|x-\mu\|^2}{2\sigma_c^2} - d\ln\sigma_c + \ln\pi_c$

↳ $P(Y=c|X) = \frac{P(X|Y=c)\pi_c}{\sum_i P(X|Y=i)\pi_i}$
$= S(Q_c(x) - \sum_{i \neq c} Q_i(x))$

LDA: Assume $x_i \sim N$ with some variance σ.

Predict $\max_c \frac{\mu_c \cdot x}{\sigma^2} - \frac{\|\mu_c\|^2}{2\sigma^2} + \ln\pi_c$

Multivariate Gaussians

$f(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu))$
↳ PSD

$\Sigma = V\Gamma V^T$

$\Sigma^{\frac{1}{2}} = V\Gamma^{\frac{1}{2}} V^T$ * Maps sphere to ellipsoid

QDA

$\hat{\Sigma}_c = \frac{1}{n_c}\sum_{i:y_i=c} (X_i - \hat{\mu}_c)(X_i - \hat{\mu}_c)^T$

Predict $\max_c -\frac{1}{2}(x-\mu_c)^T \Sigma_c^{-1}(x-\mu_c) - \frac{1}{2}\ln|\Sigma_c| + \ln\pi_c$

LDA

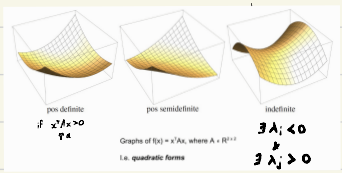
Predict $\max_c \mu_c^T \Sigma^{-1} x - \frac{1}{2}\mu_c^T \Sigma^{-1} \mu_c + \ln\pi_c$

* Likelihood of Gaussian

$L(\mu, \sigma; X_1, \ldots, X_n) = f(X_1) \cdot f(X_n)$
↳ ln, $\nabla_\mu l = 0$, $\frac{dl}{d\sigma} = 0$
$\Rightarrow \hat{\mu} = \frac{1}{n}\sum_i^n X_i$, $\hat{\sigma}^2 = \frac{1}{dn}\sum_i^n \|X_i - \mu\|^2$

For QDA: MC, Find $\hat{\mu}_c, \hat{\sigma}_c^2, \hat{\pi}_c = \frac{n_c}{\sum n_0}$

For LDA: $\hat{\sigma}^2 = \frac{1}{dn}\sum_c \sum_{i:y_i=c} \|X_i - \mu\|^2$



(b) [6 pts] Consider the two-dimensional bivariate normal distribution $N(0, \Sigma)$ where the covariance matrix Σ is the matrix you derived in part (a) and the mean is $\mu = 0$. Let $f(x)$ be the PDF of that normal distribution, where $x \in \mathbb{R}^2$. What are the lengths of the major and minor axes of the ellipse

$$f(x) = \frac{1}{4\pi\sqrt{10}}$$

Justify your answer.

The determinant of Σ is 10 (the product of the eigenvalues, or you can compute it the hard way). The normal PDF is

$$f(x) = \frac{1}{\sqrt{(2\pi)^2|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) = \frac{1}{2\pi\sqrt{10}} \exp\left(-\frac{1}{2} x^T \Sigma^{-1} x\right).$$

Setting that to $1/(4\pi\sqrt{10})$ gives

$$\begin{aligned} -\frac{1}{2} x^T \Sigma^{-1} x &= \ln\frac{1}{2} \\ \|\Sigma^{-1/2} x\|^2 &= 2\ln 2 \\ \|\Sigma^{-1/2} x\| &= \sqrt{2\ln 2}. \end{aligned}$$

The eigenvalues of $\Sigma^{-1/2}$ are $1/\sqrt{5}$ and $1/\sqrt{2}$. The major axis of the ellipse is the eigenvector x that solves this equation and has eigenvalue $1/\sqrt{5}$ (or eigenvalue 5 for Σ); thus the major axis has length $\sqrt{10}\ln 2$. The minor axis is the eigenvector x that solves this equation and has eigenvalue $1/\sqrt{2}$ (or eigenvalue 2 for Σ); thus the minor axis has length $2\sqrt{\ln 2}$.

Regression

Linear

$\min_{w,a} \sum_{i=1}^{n} (X_i \cdot w + a - y_i)^2$

- write X as $n \times (d+1)$ w/ bias
  - ↳ $\min_w \|Xw-y\|^2 = RSS(w)$

$\|Xw-y\|^2 = w^T X^T X w - 2y^T X w + y^T y$

$\nabla RSS = 2X^T X w - 2X^T y = 0$

$\Rightarrow X^T X w = X^T y$

Solve: let $w = \underbrace{(X^T X)^{-1} X^T}_{\text{pseudoinverse}} y$

↳ if invertible, unique sol

Logistic

$\min_w \sum_{i=1}^{n} L(s(X_i \cdot w), y_i)$ # $s = \frac{1}{1+e^{-y}}$, $\frac{\partial}{\partial a} S(x-a) = -S(1-S)$

$= -\sum_{i=1}^{n} (y_i \ln s(X_i \cdot w) + (1-y_i) \ln (1 - s(X_i \cdot w)))$

$J(w)$ is convex → gradient descent

↳ $\nabla J(w) = -X^T(y - s(Xw))$

↳ Can use Newton's

↳ $\nabla_w^2 J(w) = \sum_{i=1}^{n} s_i(1-s_i) X_i X_i^T$

$= X^T \Omega X$ where $\Omega = \text{diag}(s_i(1-s_i))$

$\Rightarrow w \leftarrow w + (e : X^T \Omega X e = X^T(y - s))$

Statistical Justification

- Empirical risk: $\hat{R}(h) = \frac{1}{n} \sum_i L(h(X_i), y_i)$
- Suppose $y_i \sim N(g(X_i), \sigma^2)$

  ↳ $\ell(g; X, y) = -\frac{1}{2\sigma^2} \sum (y_i - g(X_i))^2 - C$

  → estimate g by Least Square

- $\mathcal{L}(h; X, y) = \prod_{i=1}^{n} h(X_i)^{y_i} (1 - h(X_i))^{(1-y_i)}$

  ↳ $\ln \mathcal{L}(h; X, y) = -\sum L_{\text{logistic}}(h(X_i), y_i)$

  → Max likelihood → minimize logistic loss

Weighted Least-Square Regression

- Assign each x_i a weight w_i, collect w_i's in $n \times n$ diag Ω
- Find $\min_w (Xw-y)^T \Omega (Xw-y) = \sum_{i=1}^{n} w_i (X_i \cdot w - y_i)^2$

  OR solve $X^T \Omega X w = X^T \Omega y$.

Newton's Method

- At point v, approx $J(w)$ near v
- Taylor's: $\nabla J(w) = \nabla J(v) + (\nabla^2 J(v))(w-v) + O(\|w-v\|^2)$

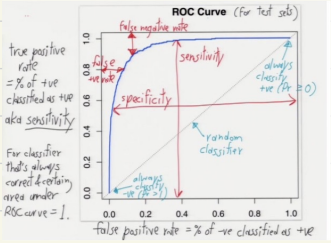
  ↳ $w = v - (\nabla^2 J(v))^{-1} \nabla J(v)$

- REPEAT: $e \leftarrow$ sol to $\nabla J^2(w)\, e = -\nabla J(w)$

  $w \leftarrow w + e$

− doesn't know if max/min, expensive Hessian, # weights, non-smooth

+ If J quadratic → 1 step, right step length to reach min.



Bias/Variance

$\text{Bias}_D[\hat{f}(x;D)] = E_D[\hat{f}(x;D)] - f(x)$

$\text{Var}_D[\hat{f}(x;D)] = E_D[\hat{f}(x;D)^2] - E_D[\hat{f}(x;D)]^2$

$$
\begin{aligned}
&E[(y-\hat{f})^2] = E[(f+\epsilon-\hat{f})^2] \\
&= E[(f+\epsilon-\hat{f}-E[\hat{f}]+E[\hat{f}])^2] \\
&= E[(f-E[\hat{f}])^2] + E[\epsilon^2] + E[(E[\hat{f}]-\hat{f})^2] + 2E[(f-E[\hat{f}])\epsilon] + 2E[(f-E[\hat{f}])(E[\hat{f}]-\hat{f})] + 2E[\epsilon(E[\hat{f}]-\hat{f})] \\
&= (f-E[\hat{f}])^2 + E[\epsilon^2] + E[(E[\hat{f}]-\hat{f})^2] + 2(f-E[\hat{f}])E[\epsilon] + 2E[\epsilon]E[E[\hat{f}]-\hat{f}] + 2E[(E[\hat{f}]-\hat{f})(f-E[\hat{f}])] \\
&= (f-E[\hat{f}])^2 + E[\epsilon^2] + E[(E[\hat{f}]-\hat{f})^2] \\
&= (f-E[\hat{f}])^2 + \text{Var}[\epsilon] + \text{Var}[\hat{f}] \\
&= \text{Bias}[\hat{f}]^2 + \text{Var}[\epsilon] + \text{Var}[\hat{f}] \\
&= \text{Bias}[\hat{f}]^2 + \sigma^2 + \text{Var}[\hat{f}]
\end{aligned}
$$

Regularization

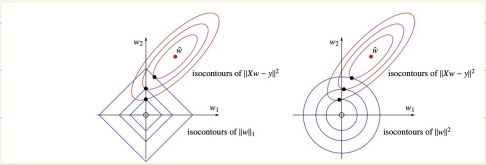
L2 (ridge)

$P(w|X,Y) = \frac{f(y|X,w) \cdot f(w)}{f(y|X)} = \frac{L(w) f(w)}{f(y|x)}$

→ $\min_w \|Xw-y\|^2 + \lambda \|w\|^2$

↳ garuntees PSD, unique sol

→ Normal eq: $(X^T X + \lambda I')\, w = X^T y$



L1 (Lasso)

→ $\min_w \|Xw-y\|^2 + \lambda \|w\|_1$ Inclined to give $w_i = 0$

Feature subset selection

All features increase var, but not all decreases bias.

Heuristic 1: Forward Step

↳ start w/ 0 features, keep adding best until err(validation) ↑

↳ $O(d^2)$

Heuristic 2: Backward step

↳ start w/ all, keep removing, $O(d^2)$