Supervised Learning

$$\begin{aligned}\text{Bayes error:} &\quad E[(y - f^*(x))^2] \\ \text{Bias:} &\quad (f^*(x) - E[A(x)])^2 \\ \text{Variance:} &\quad E[(A(x) - E[A(x)])^2]\end{aligned}$$

To derive bayes estimator, find $f^*(x)$ that minimizes $error \propto \sum_{y \in Y}(y - f*(x))^2 p(y|x)$ (take the derivative).

k-**NN:** As $m \to \infty$, 1-nn error rate is no more than twice the Bayes error (Cover's bound). k-NN optimality requires:

$$k(m) \to \infty \quad \frac{k(m)}{m} \to 0$$

A problem is **well-posed** if solution exists, it is unique and depend continuously on data. Ill-posed otherwise.

$$\begin{aligned}\textbf{Primal:} &\quad w = (X^T X + \lambda I_n)^{-1} X^T y \\ \text{Solve / predict:} &\quad O(mn^2 + n^3) \ / \ O(n) \\ \textbf{Dual:} &\quad \alpha = (XX^T + \lambda I_m)^{-1} y \\ \text{Solve / predict:} &\quad O(nm^2 + m^3) \ / \ O(m)\end{aligned}$$

Kernels are *translation invariant* if they are based on differences. Kernels are *radial* if they're t.i. AND use norms.

$$\begin{aligned}\text{Anova kernel:} &\quad K_a(x,t) = \prod_i (1 + x_i t_i) \\ \text{Polnyomial kernel:} &\quad K_p(x,t) = (1 + x^T t)^p\end{aligned}$$

For the infinite polynomial kernel:

$$\sum_{i=0}^{r} \frac{a^i}{i!}(x^T t)^i \longrightarrow e^{a\, x^T t}$$

When building a **tree** T, keep building until min. number of nodes per leaf remains. Prune on cost-complexity: $cost(T) = (\text{train error}) + \lambda|T|$.
Use Gini-index or cross entropy loss to prune (since they are sensitive), misclassification for pruning. Prune using *weakest-link* pruning.
For **bagging**, each step samples M examples from m *with replacement*. For **random forests**, each tree uses subset of k features, k usually set to $\sqrt{d}$ or $\log(d)$.

For **boosting**, if every weak learner has error $\epsilon_t \leq 0.5 - \gamma$:

$$\text{training error} \leq e^{-2T\gamma^2}$$

We use *exponential loss* for Adaboost because it does not punish positive margin and is differentiable. (Hinge is not diff., square loss punishes positive margin, misclassification is not continuous). Can think of it as forward stagewise additive model.
Bagging reduces variance, boosting reduces bias. Boosting fails with noisy data.

Non-separable SVMs use Hinge loss because misclass. is NP-hard. Optimize using slack variables ξ_i:

$$\frac{1}{2} w^T w + C \sum_{i=1}^{m} \xi_i \quad \text{s.t.} \quad y_i(w^T x_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0$$

Non-separable is same as regularization with $\lambda = \frac{1}{2C}$ with Hinge loss: $max(1 - y\hat{y}, 0)$.

$$\begin{aligned}\text{Percep:} &\quad M \leq (4k+1)(n+1) \\ \text{Winnow:} &\quad M \leq 3k(\log n + 1) + 2\end{aligned}$$

Probably (with prob. $1 - \delta$) **A**pproximately (up to accuracy ϵ) **C**orrect. For a random sample that wasn't used for training (validation set bound), with probability at least $1 - \delta$:

$$L_D(h) \leq L_V(h) + \frac{ln\frac{1}{\delta}}{2m}$$

Finite hypothesis class H is PAC-learnable with sample complexity $m_H(\epsilon, \delta) \leq \frac{\log(|H|/\delta)}{\epsilon}$. To prove $VCdim(H) = d$, need to show that one set C of size d is shattered, and that no set of size $d+1$ can be shattered.

For the graph Laplacian, $L = D - W$. First eigenvector is always 1, with eigenvalue 0.

$$\|u\|_G^2 = \langle u, u \rangle = \sum_{(i,j) \in E(G)} w_{ij}(u_i - u_j)^2$$